OPINION MINING OF UNSTRUCTURED DATA USING MACHINE LEARNING

Synopsis of the PhD Thesis

Submitted by

Ruchi Kaushal Sharma

SAP No. 70821016020

Department of Computer Engineering

Supervisor

Dr. Pravin Shrinath,

Associate Professor, MPSTME, Mumbai

submitted in partial fulfilment of the requirements of

DOCTOR OF PHILOSOPHY (Engineering)

MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT & ENGINEERING
SVKM's NARSEE MONJEE INSTITUTE OF MANAGEMENT STUDIES (Declared as Deemed-to-be University Under Section 3 of the UGC Act, 1956)
V. L. Mehta Road, Vile Parle (West)
MUMBAI – 400056

2024

TABLE OF CONTENTS

ABSTRACT iv

CHAPTER 1 1

INTRODUCTION 1

1.1 OVERVIEW 1

1.2 MOTIVATION OF THE THESIS 2

1.3 APPLICATIONS OF OPINION MINING 2

1.4 RESEARCH GAPS 3

1.5 RESEARCH OBJECTIVES (RO) 3

1.6 PROBLEM STATEMENT 3

1.7 PROPOSED METHODOLOGY 4

CHAPTER 2 7

LITERATURE REVIEW 7

2.1 OVERVIEW 7

2.2 OPINION MINING TAXONOMY 7

2.3 PHASES OF OPINION MINING FRAMEWORK 8

2.4 OPINION MINING RESEARCH GAPS 9

CHAPTER 3 10

DESIGN AND DEVELOPMENT OF PROPOSED RHIEN CORPUS FOR LOW RESOURCE AVAILABILITY 10

3.1 OVERVIEW 10

3.2 PROPOSED METHODOLOGY 10

3.3 EXPERIMENTAL RESULTS 13

3.4 FINDINGS AND DISCUSSION 15

CHAPTER 4 17

WEIGHTED CODE MIXED OPINION MINING FRAMEWORK (WCM-OMF) 17

4.1 OVERVIEW 17

4.2 DATASET STATISTICS 17

4.3 PROPOSED OPINION MINING FRAMEWORK 17

4.4 WORKING MECHANISM OF WCM-OMF METHOD 19

4.5 EXPERIMENTAL RESULTS AND DISCUSSION 21

CHAPTER 5 23
FENNEC FOX OPTIMIZATION OPINION MINING FRAMEWORK (FFO-OMF) USING HYBRID APPROACH 23
5.1 OVERVIEW 23
5.2 PROPOSED METHDOLOGY 23
5.3 EXPERIMENTAL RESULTS AND DISCUSSION 26
CHAPTER 6 27
ADAPTIVE FENNEC FOX OPTIMIZATION OPINION MINING FRAMEWORK 27
6.1 BACKGROUND 27
6.2 PROPOSED METHDOLOGY 27
6.3 EXPERIMENTAL RESULTS 28
6.4 FINDINGS AND DISCUSSION 30
CHAPTER 7 32
ENSEMBLE CODE MIXED OPINION MINING FRAMEWORK (ENCOMF) 32
7.1 OVERVIEW 32
7.2 PROPOSED METHDOLOGY 32
7.3 EXPERIMENTAL RESULTS 34
7.4 FINDINGS AND DISCUSSION 36
CHAPTER 8 37
CONCLUSION AND FUTURE WORK 37
8.1 CONCLUSION 37
8.2 CONTRIBUTION SUMMARY 38
8.3 FUTURE WORK 38
LIST OF PUBLICATIONS 39
REFERENCES 40

ABSTRACT

Keywords: Opinion Mining; Unstructured Data; Machine Learning; Code-Mixed Analysis

Opinions are a crucial source of information in enterprises for Business Intelligence (BI). They have the potential to be beneficial in several applications, including bioinformatics, finance government sector. However, these applications cannot utilize the digital opinions available online directly due to their unstructured format, thereby rendering them unusable for decision-making purposes. There are specific challenges to Opinion Mining of Unstructured Data. This data is in the form of free-form text with inconsistent lexical structure, non-colloquial, shortened length, and sometimes ambiguous. Moreover, these opinions are expressed as a combination of several units of lexicons within a singular utterance. In our research work, we analyse these challenges and overcome the limitations found with the existing system to a significant extent. The issues to handle the problem of unstructured data and code mixed opinions are addressed, and opinion mining frameworks are implemented.

In this thesis, we present the design and development of a curated corpus to solve the problem of Low Resource Availability (LRA) in unstructured data. Further, to eliminate the problem of Code Switching (CS), we designed and implemented an adaptive weighted lemma that is used as input for the proposed "Weighted Code Mixed Opinion Mining Framework (WCM-OMF)". The performance results of the proposed opinion mining system are presented in the form of accuracy, precision, recall, and F1 scores in the thesis and compared with those of existing opinion mining systems. The proposed WCM-OMF framework for unstructured data presented in our research is found to be more effective than existing opinion detection frameworks.

Furthermore, the design of the proposed "Fennec Fox Optimization Based Opinion Mining Framework (FFO-OMF)" and "Adaptive Fennec Fox Optimization Opinion Mining Framework (AFFO-OMF) are implemented. AFFO-OMF is found to be the best-performing model among the proposed solutions, achieving an accuracy of **88.9%** and a precision of **89.32%.** Additionally, we have employed linguistic metrics for the qualitative validation of the proposed corpus as efficient input for opinion classification frameworks. Further, we implement benchmarking of our proposed corpus using standard Machine Learning algorithms, develop an Ensemble Code Mixed Opinion Mining Framework (ENCOMF). Results obtained from proposed opinion mining frameworks are validated using both qualitative and quantitative metrics. Out of the 3 proposed opinion mining frameworks, AFFO-OMF is found to be the most effective considering the standardized performance parameters on benchmark dataset.

CHAPTER 1

INTRODUCTION

This thesis investigates the challenges of online opinion mining, focusing on opinion detection and classification. It aims for the design and development of a comprehensive opinion mining framework for efficient opinion classification. Further, the opinion mining framework developed can classify opinions for a combination of more than one lexicon. In recent times, there has been an exponential increase in data available online in the form of opinions present in blogs, forums, and other online platforms. Social Media Intelligence (SMI) research domain utilizes online opinion data to extract meaningful information for efficient decision making.

1.1 OVERVIEW

Opinion mining (OM) is a subfield of Natural Language Processing (NLP) and is the systematic extraction and computational analysis of the opinion. Although, there is an exponential volume of user-generated content available online for social informatics [1], it is primarily in the unstructured format. [2]. Code mixing is alternating between two or more languages in a single opinion text source, rendering it complex to implement and interpret [3]. Fig. 1.1 depicts the various opinion mining challenges and related concepts.

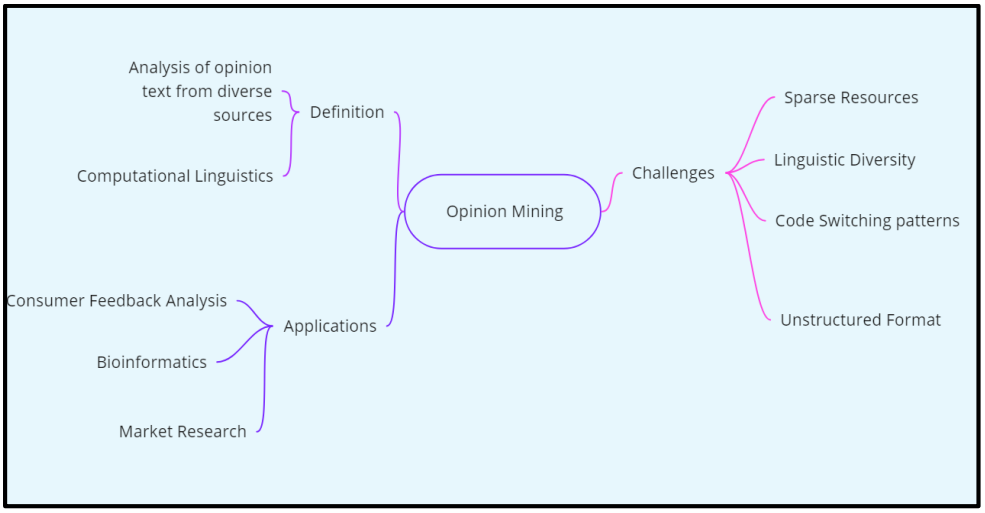


Fig. 1.1 Opinion Mining Concept, Applications, and Challenges

As per recent surveys by MIT, eighty percent of the online data is heterogeneous, in an unstructured format, and has a non-standardized dialect [9]. Classifying this text-based opinion is challenging despite its unstructured format. Enterprises with a significant online presence benefit from this opinion detection framework for actionable insights to achieve business intelligence (B.I.) [5].

1.2 MOTIVATION OF THE THESIS

User-generated opinions on blogs, social media, and review sites have grown exponentially, providing an increasing amount of unstructured data. An end-to-end framework is required to interpret unstructured opinions, thus improving overall efficiency. These opinions are expressed in a medium convenient to them , are bilingual or multilingual and challenging to pre-process. Further, opinions are short, non-colloquial, and have unstructured formats that are complex to interpret[6]. This research thesis focuses on opinion mining expressed in an unstructured format.

1.3 APPLICATIONS OF OPINION MINING

Despite the massive amount of online opinions, it has limited value since it is in an unstructured format, rendering it unusable. To overcome this problem, Opinion Mining, also termed as sentiment analysis, has recently evolved into a critical research domain.

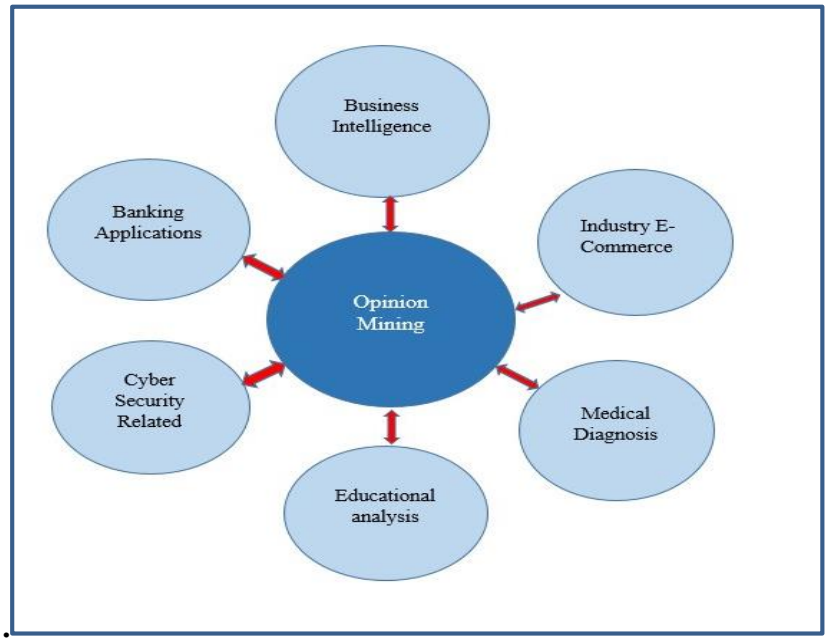


Fig. 1.2: Opinion Mining Application Domains [7]

Fig. 1.2 explains the application domain of the proposed research in detail.

Government Applications: Opinion mining is useful in the governmental sector to gauge public opinion for policy reform, general governance, natural disasters, pandemics[5][8] .

Finance: Opinion Mining frameworks would aid organizations in the financial domain in analyzing market stock, shareholdings, or other financial holdings [9].

Healthcare and Bioinformatics: Opinion mining can be effectively used in the healthcare sector to identify public health trends and enhance healthcare diagnosis [10].

Enterprise Project Management: Feedback from client firms and beta users is beneficial for project leaders to achieve successful project completion [11].

1.4 RESEARCH GAPS

Through extensive analysis, we derive the following research gaps stated below:

a. There is a lack of annotated text corpora for unstructured Opinion Mining, leading to Low Resource Availability (LRA) issues [12].
b. Due to unstructured components in opinions, the input corpus lacks adherence to a standardized format and involves complex processing [13]. The challenge related to unstructured opinion mining is not fully addressed.
c. Lack of development of effective opinion code-switching methodologies [14].
d. Lack of efficient feature selection technique for opinion detection [12].
e. Focus on only one aspect for developing an opinion mining framework, considering qualitative or quantitative aspects in isolation. [12] [15].

1.5 RESEARCH OBJECTIVES (RO)

RO1: **To Design and Develop Benchmark Dataset for Code Mixed Opinions having Low Resource Availability (LRA):** RO1 aims to provide a benchmark dataset due to the challenge of LRA in code-mixed opinions. Opinion mining research in unstructured opinions using Machine Learning techniques requires a massive volume of data for training.

RO2: To Handle Non-Colloquial Linguistic Variability in Unstructured Data: RO2 develops strategies to manage unstructured data's various linguistic features, especially code-mixed opinions. These issues are addressed by the pre-processing model approaches developed toward this objective.

RO3: To Design and Implement Code-Switching Opinion Mining Model: RO3 aims to implement an opinion mining model that handles code switching, where speakers use multiple lexicons in a singular utterance. Traditional opinion mining techniques perform inefficiently with code switching since it complicates the textual component. The proposed method enables the model to interpret the code-mixed opinion text for user-generated content interpretation correctly.

RO4 Opinion Detection Feature Selection Algorithm Design and Implementation: RO4 designs and implements a feature selection technique for code-mixed text opinion detection, as traditional approaches lack efficiency due to the diversity and complexity of code-mixed languages. The goal is to implement a hybrid method using statistical and linguistic analysis for feature selection.

RO5: To implement a Hybrid Opinion Mining Framework Development : It is implemented using machine learning methodologies and natural language processing in an Ensemble approach.

1.6 PROBLEM STATEMENT

"To design and implement Opinion Mining framework for Unstructured Data using Machine Learning"

1.7 PROPOSED METHODOLOGY

In this section, we provide an overview of the proposed methodology. Fig. 1.3 and Fig. 1.4 depict the differences between the Generalized Opinion mining framework and the Proposed Framework.

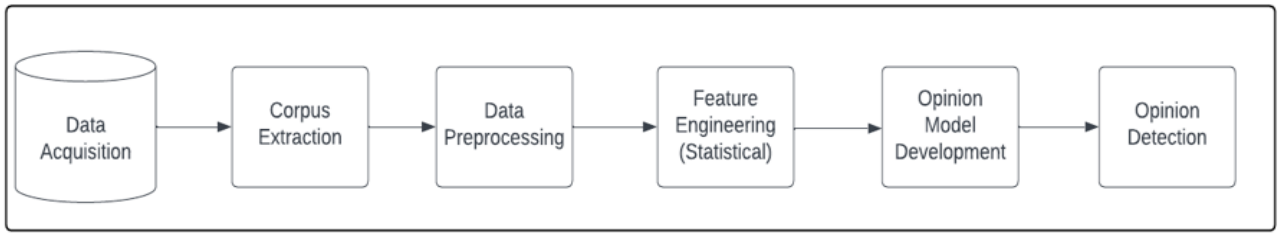


Fig. 1.3: Generalized Opinion Mining Framework [25]

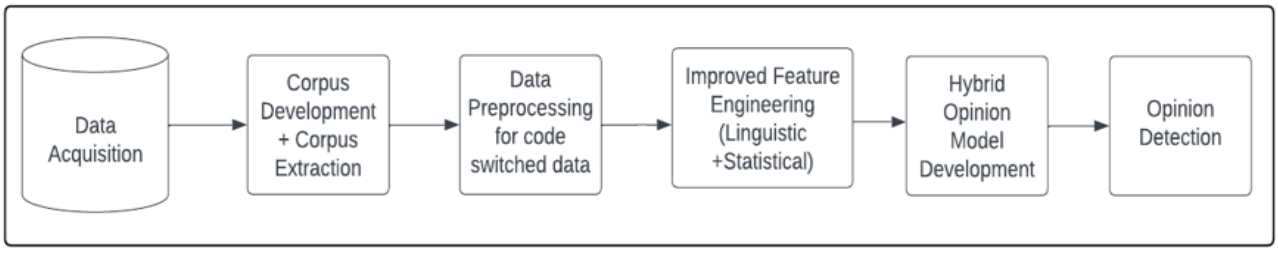
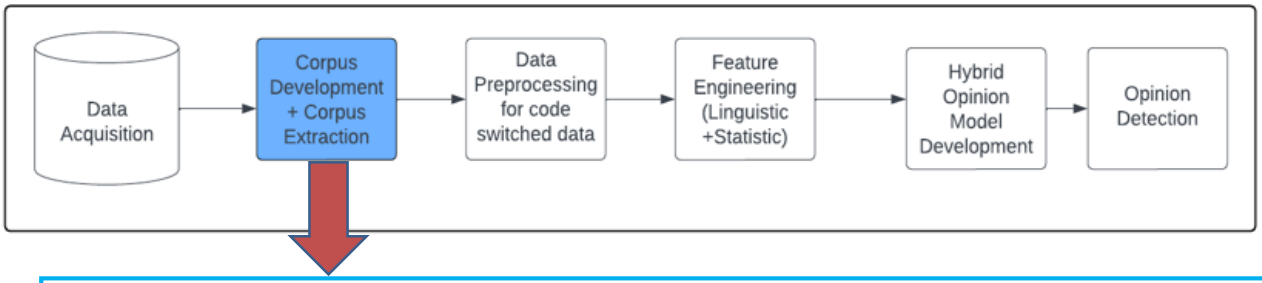

Data Acquisition
Corpus Development + Corpus Extraction
Data Preprocessing for code switched data
Improved Feature Engineering (Linguistic +Statistical)
Hybrid Opinion Model Development
Opinion Detection

Fig. 1.4: Proposed Opinion Mining Framework

PHASE I OF PROPOSED SOLUTION

Fig. 1.5 illustrates the proposed corpus Romanized Hindi and English mixed opinions (RHIEN) developed, which is a significant contribution to the domain of opinion mining for analysis of code-mixed opinions.



a. Development of proposed corpus RHIEN (Romanized Hindi and English mixed opinions)
b. Proposed solution implements unstructured opinion handling.

Fig. 1.5: Phase I of Proposed Opinion Mining Framework

PHASE II OF PROPOSED SOLUTION

Phase II of the proposed solution described in Fig. 1.6 includes generating stop words suitable for code-mixed languages that cannot use traditional stop word lists since they do not consider linguistic variation in opinions.

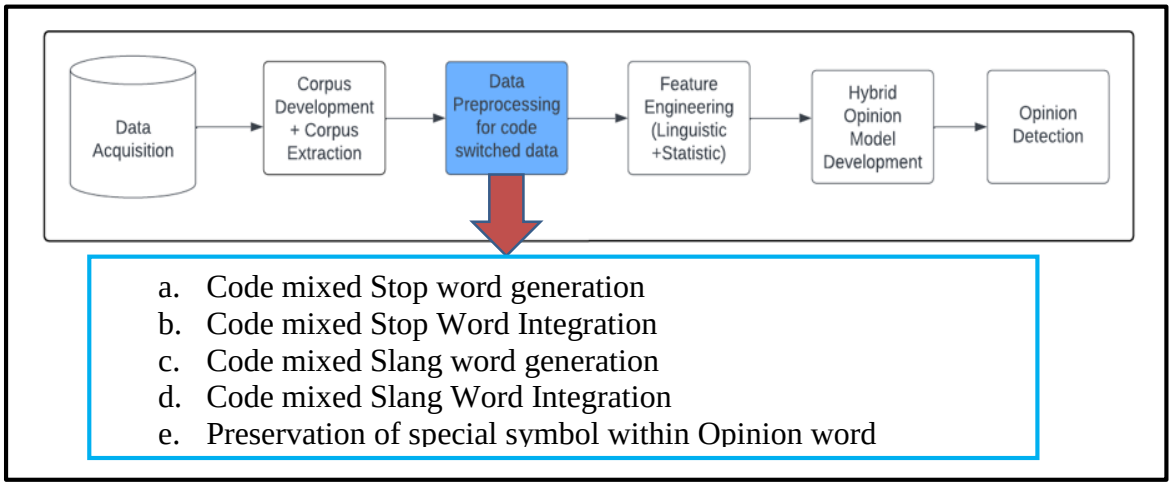

Data Acquisition
Corpus Development + Corpus Extraction
Data Preprocessing for code switched data
Feature Engineering (Linguistic +Statistic)
Hybrid Opinion Model Development
Opinion Detection
a. Code mixed Stop word generation
b. Code mixed Stop Word Integration
c. Code mixed Slang word generation
d. Code mixed Slang Word Integration
e. Preservation of special symbol within Opinion word

Fig. 1.6: Phase II of Proposed Opinion Mining Framework

a. **Generation of Code-Mixed Stop Words:** In this phase, we generate an extensive list of frequently used stop words that can be identified from codemixed texts across different language backgrounds, thereby improving the overall performance of opinion mining models by removing non-informative terms more efficiently.
b. **Integration of Code-Mixed Stop Words:** The generated codemixed stop words are integrated into the opinion mining pipeline. This integration ensures that the pre-processing step improves machine learning models by removing noise and concentrating on only lexical features conveying opinions for opinion classification in code-mixed settings.
c. **Slang Word Generation in Code-Mixing:** This phase implements a slang dictionary mapping programming function for the linguistic combinations to overcome the issue of slang in unstructured data.
d. **Slang Word Integration in Code-Mixing:** The slang word utterances are mapped with the opinion corpus and substituted using the key-value pair method based programming implementation, ensuring accurate representation through a mapping approach with the corresponding opinion in the corpus.

PHASE III OF PROPOSED SOLUTION

Fig. 1.7 depicts Phase III of the proposed solution related research objective of Feature Engineering.

a. The proposed WCM-OMF framework implements a weighting mechanism for linguistic features to enhance opinion mining in code-mixed data for nuanced opinion classification.
b. The design and implementation of the proposed FFO-OMF framework presents a significant contribution to the field of Opinion Mining implements the hybridization of the Fennec Fox Optimization algorithm and Natural Language Processing methodologies for unstructured opinion mining.

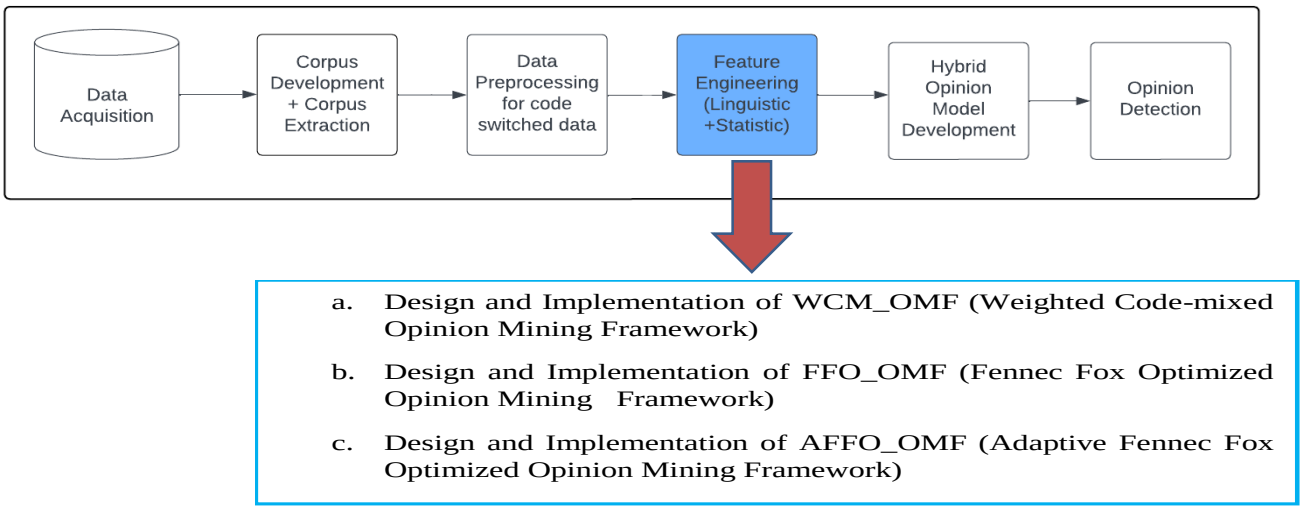

Data Acquisition
Corpus Development + Corpus Extraction
Data Preprocessing for code switched data
Feature Engineering (Linguistic +Statistic)
Hybrid Opinion Model Development
Opinion Detection
a. Design and Implementation of WCM_OMF (Weighted Code-mixed Opinion Mining Framework)
b. Design and Implementation of FFO_OMF (Fennec Fox Optimized Opinion Mining Framework)
c. Design and Implementation of AFFO_OMF (Adaptive Fennec Fox Optimized Opinion Mining Framework)

Fig. 1.7: Phase III of Proposed Opinion Mining Framework

c. The proposed AFFO-OMF solution is an extension of FFO-OMF, which is the basis of the optimization process that enhances FFO model performance. The framework presents a modified approach of FFO-OMF based on the linguistic features of the input data, thereby ensuring optimal performance in various contexts.

PHASE IV OF PROPOSED SOLUTION

Fig. 1.8 depicts Phase IV of the proposed solution related to the research objective for Hybrid Opinion Model development and corresponding opinion detection.

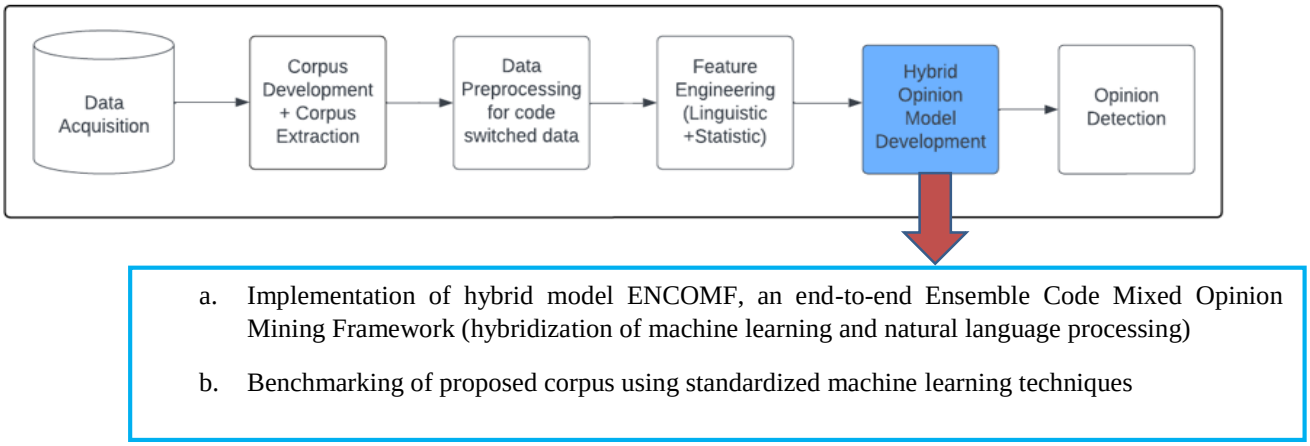

Data Acquisition
Corpus Development + Corpus Extraction
Data Preprocessing for code switched data
Feature Engineering (Linguistic +Statistic)
Hybrid Opinion Model Development
Opinion Detection
a. Implementation of hybrid model ENCOMF, an end-to-end Ensemble Code Mixed Opinion Mining Framework (hybridization of machine learning and natural language processing)
b. Benchmarking of proposed corpus using standardized machine learning techniques

Fig. 1.8 Phase IV of Proposed Opinion Mining Framework

The ENCOMF (Ensemble Code Mixed Opinion Mining Framework) proposed solution is a significant development for unstructured opinion mining. This hybrid architecture combines machine learning and natural language processing (NLP) techniques to create an end-to-end approach for opinion mining in code-mixed languages. It addresses unstructured data mining and benchmarking against established baselines providing efficient opinion mining model using standardized machine learning and embedding techniques.

CHAPTER 2

LITERATURE REVIEW

2.1 OVERVIEW

The previous chapter discussed the fundamental concepts, scope, problem statement, research objective, and proposed methodology for the Opinion Mining Framework. This chapter presents the recent techniques, advantages, and limitations of the existing Opinion Mining systems. Further, opinion mining taxonomy, comparative study of existing research, and open research areas are analysed.

2.2 OPINION MINING TAXONOMY

Supervised Opinion Mining: In this method, the researchers adopt conventional machine learning techniques, namely Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression. Opinion mining studies use ML [16] for various domains, namely the work done on the climate change crisis, which highlights the priority levels of regional dialects [17].

Unsupervised Opinion Mining: It is observed that authors use unsupervised opinion mining in various application domains. For instance, medical domain analysis and deep clustering with embedding was used for recommendation using document vectors. In further studies, deep learning techniques have been applied with promising results [18]. Other studies used deep learning-based methods [19][20] and stacked layering mechanism.

Hybrid Opinion Mining: Various approaches are adopted with pre-processing and classification combination variations for hybridization-based opinion mining. In [21], authors use sentence-level embedding combined with the Transformer approach. In [22], authors hybridize bio-inspired optimization algorithms with neural networks (RNN). Other combinations include neural network and transformer approaches [23].

Researchers [24] achieve good performance employing traditional J48 machine learning approach on self-developed data of content extracted. The research study [25] implements deep learning and dual datasets for cross-pair mapping purposes. In [26], the authors highlight the lack of a training corpus for Dravidian language content classification into multiple classes. The target entity extraction was successful through decisions based on neighboring information but limited by CNN capabilities. The shallow neural network is for Spanish input and achieves an F1 score of 0.48 [27].

Table 2.1 Comparative Analysis of Result Findings and corresponding methods of recent state-of-the-art Opinion Mining Systems

Ref.	Author and Publisher	Year	Opinion Detection Method	Outcomes	Limitations
[32]	Baruah et al., IIITG-ADBU	2022	BERT classifiers, SVM classifiers	0.67 F1 score achieved	Domain dependence
[33]	SYadav, ACL	2022	Neural Network	Achieved an F1-score of 0.67 using LSTM	Code switching data challenge
[35]	Banerjee, S., et al.	2020	Recurrent Neural Network and Convolution Neural Network	F1 score of 0.69	Error analysis reveals data ambiguity due to an unstructured format
[36]	Shekhawat, S. S., et al. Evolutionary Intelligence, Springer	2021	Hybrid Spider Monkey Optimization	Accuracy of 88.93	The input dataset is single opinion language. Only unstructured data is considered, not code mixing.
[37]	Dowlagara S, et al., IIITH, ACL	2021	FastText and Transformer method	F1 score of 0.66 and accuracy of 67%	Lack of training data caused misclassification errors.
[38]	Singh, K.S. Manoj, Journal of Adv. Comput.Sci.	2017	Various machine learning methods surveyed	Social opinions are difficult to interpret directly	Data sparsity Challenges and unstructured data research areas not thoroughly addressed

Table 2.1. presents the comparative analysis of the literature review findings. In [28], the authors pre-processed the data and generated the vectorized corpus using pre-trained model. Studies [27][29] utilize variations of other deep learning techniques to classify online content. The various contexts of words are handled using a stacked word embedding implementation . Discussion of research for offensive text in challenging environments is highlighted [30]. Machine learning techniques are widely adopted, followed by deep learning and hybrid techniques [31][34].

2.3 PHASES OF OPINION MINING FRAMEWORK

The majority of the existing frameworks rely solely on a single lexicon or medium, resulting in a significant loss of relevant information. Fig 2.1 describes the OM phases in detail:

a. Data extraction and retrieval: This phase retrieves online content from heterogeneous web sources and unstructured data platforms using programming APIs.
b. Opinion Preprocessing: In the next phase, subtasks of NLP, including Stemming, Lemmatization, and irrelevant noise removal, are executed.
c. Lexicon Detection: The phase for detection of lexicon programmatically is crucial for efficient machine learning and deep learning model implementation.

d. Translation and transliteration: This phase converts from 1 lingual and structural format to another, enabling efficient heterogeneous opinions detection.
e. Feature Engineering: It comprises Feature extraction, transformation, and subsequent vectorization executed employing NLP into several diversified opinion classes.
f. Evaluation: Performance indicators vary depending upon subtasks for measuring the effectiveness of training and test evaluation.

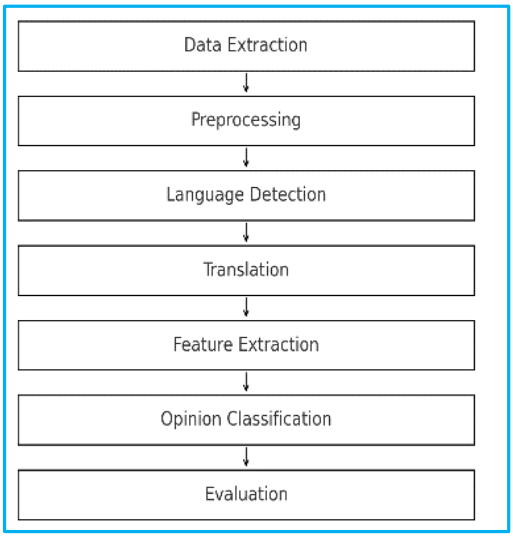


Fig. 2.1: Methodology of Opinion Mining Framework

As observed in the thematic research area map developed, "Multilingualism" and “code-mixing” in OM are still an underexplored subdomain.

2.4 OPINION MINING RESEARCH GAPS

This section presents open research issues after thoroughly investigating recent research work highlighted in detail.

a. High-quality corpus, curated to the problem statement, is an open research area and data quality is significant in a Machine Learning model for Opinion Mining subfield.
b. The output of heterogeneous slang opinions translation process, especially phrases, can lower the machine learning model's performance. Linguistic phrase dictionaries require accurate knowledge mapping systems for this dual mapping and context variations.
c. Code-switching: Interpreting opinions in mixed-language opinions is complex to implement due to code-mixing and further work is required for complex lexicon interpretation.
d. The development of a domain-independent ML model to interpret unstructured opinions is an open research avenue for practitioners. Unique lexicon, structure, vocabulary of opinions must be handled efficiently. Existing Opinion mining systems do not entirely overcome the challenges of code mixed and unstructured opinions.

CHAPTER 3

DESIGN AND DEVELOPMENT OF PROPOSED RHIEN CORPUS FOR LOW RESOURCE AVAILABILITY

3.1 INTRODUCTION

The literature review emphasized the lack of high-quality, annotated datasets that are comprehensive and representative of linguistic variations. To address this research gap, the first objective of the proposed approach is the development of a corpus(dataset) specifically curated to the research goal of unstructured opinions available online. The contributions of this chapter are summarized below:

a. Design and development of a curated corpus - RHIEN - Romanized Hindi and English code-mixed corpora
b. Code mixed Slang word generation and integration
c. Codemixed Stop word generation and integration
d. Preservation of Special characters within word w(i) in O(i) to prevent loss of information where i : i^{th} Opinion in opinion corpus where $0 \leq i \leq n$ and n: Total no of Opinions

RHIEN contributes to the domain of code-mixed opinion analysis to provide researchers and enterprises with a valuable resource to analyze and train code-mixed opinions. The corpus reflects the unstructured data and lexicon features of code-mixed language varieties often found in heterogeneous digital communication platforms.

3.2 PROPOSED METHODOLOGY

There exist preprocessing challenges of opinions extracted from online heterogeneous sources. Fig. 3.1 depicts the proposed contributions of this chapter explained in detail below. Here, the Yellow highlighted cells in the above figure represent contribution.

Code Mixed Slang Word Generation and Integration: Slang words, often informal, and comprising of heterogeneous utterances are handled that enhances the classification process during the analysis of user generated content.

Codemixed Stop Word Generation and Integration (200 Stop Words Generated): The proposed contribution consists of generating a 200 unique stop word list for the code-mixed language pair. In addition, we incorporate this generated list into the pre-processing pipeline, to deal with lexical noise and enhance the quality of the opinion mining models as depicted in Fig. 3.2.

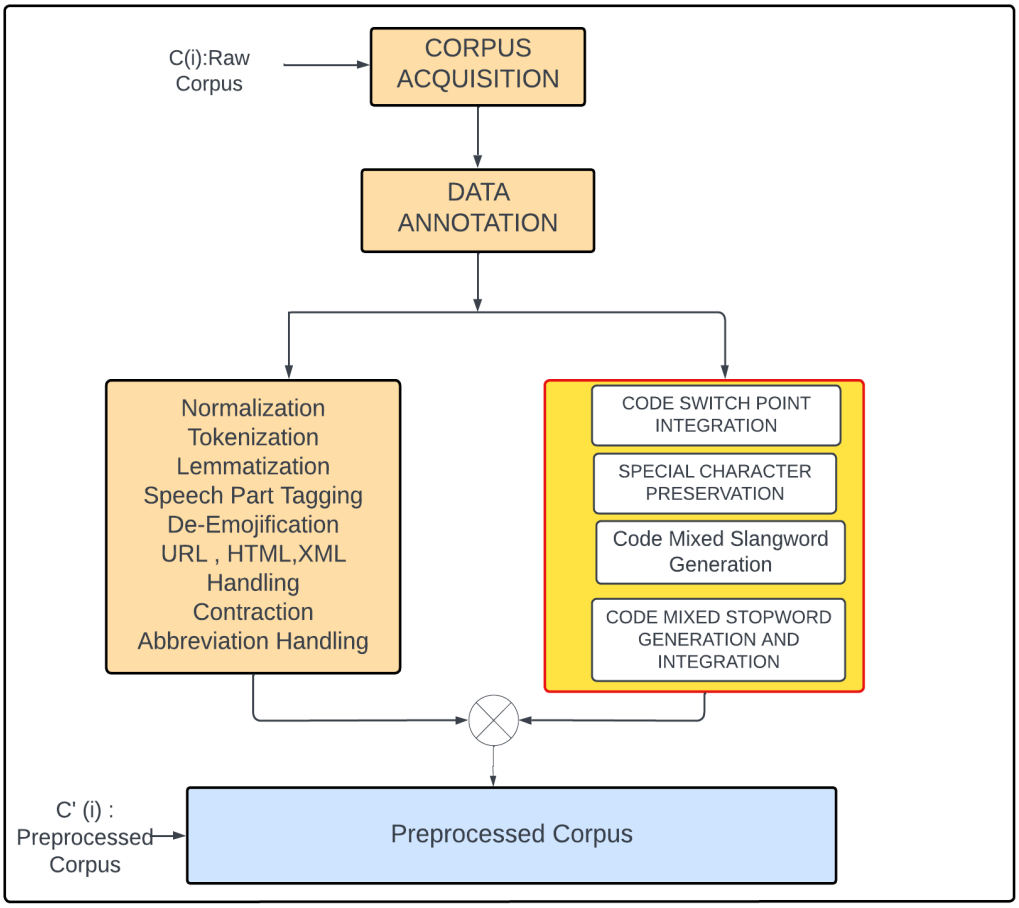


Fig. 3.1: Workflow of the Proposed Corpus Development

Preservation of Special Characters Within Word Utterances to Prevent Loss of Information: During preprocessing, we maintain special word-form characters to retain information quality in a code-mixed corpus. Special characters often span syntactic or meaning functions for non-colloquial use of expression. The proposed function ensures that special characters at the word level are retained to prevent information loss during text normalization's cleaning and tokenization phase.

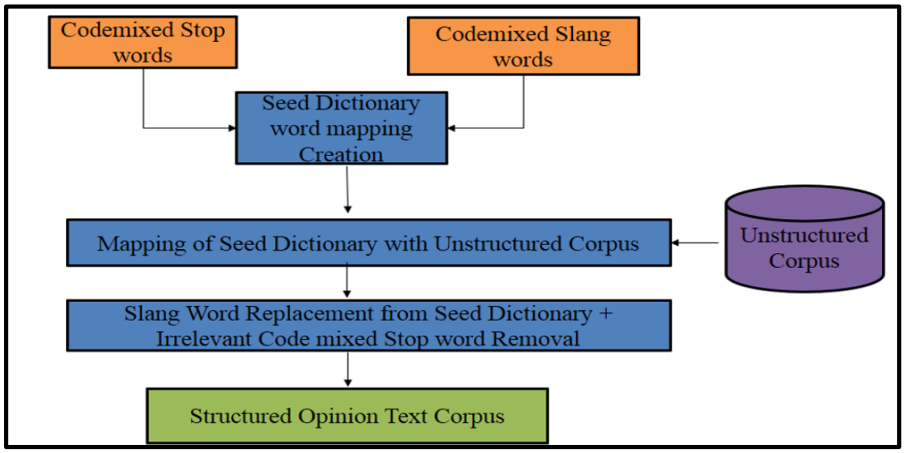

Codemixed Stop words
Codemixed Slang words
Seed Dictionary word mapping Creation
Mapping of Seed Dictionary with Unstructured Corpus
Unstructured Corpus
Slang Word Replacement from Seed Dictionary + Irrelevant Code mixed Stop word Removal
Structured Opinion Text Corpus

Fig. 3.2: Proposed Preprocessing module for Code Mixed Unstructured Data handling

Table 3.1: Instances of Unstructured Opinions Before Processing and Post-Processing Phase in the Existing Systems

Challenges of unstructured opinions(opinion Diversity)	Explanation	Example of Opinion O(i) Before Processing	Example of Opinion O(i) post processing
Punctuation Removal	Removing all punctuation marks from the text.	Wow, this new airport is amazing; exactly what I needed!	Wow this new airport is amazing exactly what I needed
Whitespace Removal	Eliminating extra spaces, tabs, or newline characters.	The service at the restaurant was exceptional.	The service at the restaurant was exceptional.
Spell Correction	Correcting spelling errors in the text.	I absolutly love this app!	I absolutely love this app!
Removing Numbers	Removing digits and numbers from the text.	I have visited this place 5 times in 2020.	I have visited this place times in .
Removing Emojis and Special Characters	Eliminating emojis, symbols, and any non-alphabetic characters.	😂 This is hilarious!!! 😂	This is hilarious
Hashtag Removal	Removing hashtags from the text.	Loved the talk by @speaker #inspiring #tech	Loved the talk by @speaker
URL Removal	Removing URLs from the text.	Check out this website for more info: https://example.com	Check out this website for more info:
Email Address Removal	Removing email addresses from the text.	Any questions, email us at contact@example.com	Any questions, email us at
HTML Tag Removal	Removing HTML tags from the text for clean text extraction.	This is a bold move!	This is a bold move!
Converting to ASCII	Converting characters to their ASCII equivalent to standardize encoding.	Café becomes Cafe	Cafe becomes Cafe

Table 3.2: Instances of Unstructured Opinions Before Processing and Post-Processing Phase for Unstructured and Code-Mixed Opinions

Challenges of unstructured opinions(opinion Diversity)	Explanation	Example of Opinion O(i) Before Processing	Example of Opinion O(i) post processing
Code-Switch Point Detection	Identifying points in the text where the language switches from Hindi to English or vice versa.	OMG, ye khana toh next level ka tasty hai!	OMG, ये खाना तो next level का tasty है!
Slang Word Expansion	Expanding slang words to their standard English or Hindi equivalents.	Yeh movie ekdum lit thi, bhai.	This movie was extremely good, brother.
Abbreviation	Expanding abbreviations to their full form for clarity.	BTW, yeh movie ekdum top thi, IMHO.	By the way, yeh movie ekdum top thi, in my humble opinion.
Contraction	Expanding contractions to their full form to improve understanding.	Can't believe it's already over, tho it was gr8.	Cannot believe it is already over, though it was great.
Segmentation Transliteration to Standard Script	Converting Hinglish text written in Roman script to Devanagari or vice versa to maintain consistency.	Ye movie bhot hi badhiya thi, totally worth it!	[ये मूवी बहुत ही बढ़िया थी], [totally worth it]!
Preservation of certain special characters	Preserving @ symbol remains intact within a Hinglish word when it appears between two	Follow k@ro guys for more updates, really worth it!	Follow karo guys for more updates, really worth it!
Code mixed Stop Words Removal	Removing Hinglish stopwords to focus on meaningful words.	Seriously, yah movie toh ekdum amazing thi, storyline bhi kamaal ki thi.	Seriously, movie ekdum amazing thi, storyline bhi kamaal ki thi.

Tables 3.1 and 3.2 illustrate the diversity present in unstructured opinions and corpus processing techniques implemented.

9 Iterations Performed for Preprocessing: In total, nine iterative processing iterations were implemented to improve the input data for issues with slang integrated usage, removing stop words, and special character usage to aid in the improvement of the corpus at the word level.

3.3 EXPERIMENTAL RESULTS

This section provides a detailed experimental result for the developed corpora. Table 3.3 provides the qualitative evaluation through linguistic metrics for the developed corpus.

Table 3.3: Qualitative Metric Evaluation of Proposed RHIEN Corpus

Linguistic Metric	Value
Number of Tokens	178,236 tokens
Number of Opinions Text	8,889
Average Opinion Text Length	112.45 characters
Vocabulary Size	24,630 unique words
Opinion Length Distribution	Min: 1
	25th Percentile: 46
	75th Percentile: 166
	Max: 1,437 characters
Number of Sentences per Opinion	1.0 sentences per opinion
Average Word Length	4.56 characters

Opinions Volume: The proposed corpus consists of large text corpora of 8,889 opinions for training machine learning models requiring adequate corpus size. The context and the size of the dataset ensure unstructured data variability. Furthermore, incorporating diverse opinions enhances the robustness for generalizable models.

Proposed Corpus Diversity (24,630 Unique Words) - Ensures Nuanced Opinion Capture: Diversity is present throughout the corpus, comprising 24,630 unique words. For unstructured opinion mining research objective, having a diverse vocabulary enables the model to detect the various linguistic hedges and nuances.

Opinion brevity (Avg. Length: 112.45): The average word count of the opinions in the corpus is 112.45, which introduces additional complexity in model development. Shorter texts tend to have less context and with possibility of ambiguous language thereby making it difficult to infer class in shorter texts. The challenge is developing an opinion mining model that efficiently conveys its accurate class for shorter texts.

Opinion Expressiveness: The proposed corpus shows the expressiveness of extracted opinions, with qualitative evaluation highlighting the complexity of opinion classification. This generalization offers framework stability for predicting expressiveness in fewer words.

Unstructured Nature: The unstructured nature is especially complex due to its varied nature, for instance, slang usage, mixed utterances, and grammatical incoherence. The metric evaluation process and its statistical significance are discussed below:

a. Flesch Score

Flesh Score indicates sentence complexity [39]. It is used to assess the readability and complexity of code-mixed text. High scores imply uncomplicated text, whereas low scores suggest a complex topic. Here, an FKS score of 2.39 (a low score) indicates the presence of complex, unstructured data, making it suitable for input to the opinion mining model.

$$Flesch\ score\ =\ 206.835 - (1.015\ x\ ASL) - (84.6\ x\ ASW) \quad (2.1)$$

Where: ASL = Average Sentence Length; ASW = Average syllables / word

$$\text{Flesch Kincaid}\ = 206.835\ - 1.015\ *\ \frac{W_i}{N(S)} - 1.015\ *\ \frac{Sy_i}{W_i} \quad (2.2)$$

Where: W_i : Number of words where $0 \leq i \leq n$;

N(S): Number of sentences in source Corpus Sy_i : Opinion Syllables .Table 3.4 explains the linguistic validation for a sample opinion instance and qualitative metric results.

Table 3.4: Performance Results of Proposed RHIEN Corpus for Sample Opinion

Sr No	PROPOSED CORPUS RHIEN DESCRIPTION	
	Qualitative metric	**Value**
1	Sample Opinion ID	6403
2	Tokenized column	Neend kaise aati hai this phaila kar chain se sojata ho ya bechain rahta ho.Had Hoti hai brother boss
3	Flesch score	103.687
4	Dale-Chall Score	18.235
5	Sentence count	2
6	Syllable count	21
7	No of characters	85
8	Flesch-Kincaid Score	2.39

b. Dale Chall

$$\boldsymbol{Dale - Chall}\ =\ 0.1579\ *\ (PDW) + 0.0496\ *\ ASL \quad (2.3)$$

Where: PDW: Percentage of Difficult Words; ASL: Average Sentence Length [40]

c. Average Number of Syllables Per Word (ASW):

To assess code-mixed content's readability and interpretability, ASW is used to indicate a complicated language mixture or advanced language [41]. Understanding ASW improves linguistic models for code-mixed text components in opinion mining. For example, in the proposed corpus, the ASW for opinion ASW[O(I)] is 21, indicating a higher complexity level input, making it difficult to interpret directly, where O(I) is the ith opinion in the opinion corpus, $0 \leq i \leq n$; here, n signifies the total number of opinions. The pre-processing functions implemented on the proposed corpus include decoding HTML, removal of emoticons, mentions (usernames), URL links, case folding, contraction expansion, handling of hashtags, punctuation marks, numbers, and special characters, removal of extra white spaces, duplicate opinion removal, and removal of null values.

3.4 FINDINGS AND DISCUSSION

The discussion and findings from experimental results are presented below:

Handling Limited Resource Challenge: In this phase, the LRA challenge is handled by developing a specialized corpus by extracting social media opinions from publicly accessible social media sites via programming APIs. Next, these opinions were then annotated by two annotators, and an inter-annotator agreement (ITA) score [42] using Cohen Kappa was computed.

$$\text{Cohen Kappa score } \kappa = \frac{\{p_o - p_e\}}{\{1 - p_e\}} \quad (2.4)$$

Where: p_o is the relative observed agreement among raters; p_e is the expected agreement and κ measures the agreement between the raters. ITA score is 94%, indicating reliable annotation for qualitative text-based opinion annotation indicating qualitative reliability of the corpus to solve LRA.

Slang Word Mapping and Substitution: The proposed methodology handles slang words appearing in code-mixed text by substitution with standardized words from a seed dictionary implemented. The substitution function minimizes corpus noise, addressing lexicon variability.

Preprocessing for Opinion Structuring: The proposed method reduces code-mixed stop words mapped to the seed dictionary for opinion structuring by reducing noise or redundancies in the text structured opinions, leading to better inference.

Structured Opinion Generation: The proposed methodology creates structured opinion representations through linguistic analysis to process this data coherently.

Evaluation of Data Structuring Efficacy: The proposed function evaluates the effectiveness of structured opinions based on the linguistic readability metric of the transformed opinions as input for the next phase of the opinion mining pipeline. Fig 3.3 shows the frequency distribution of the developed corpora.

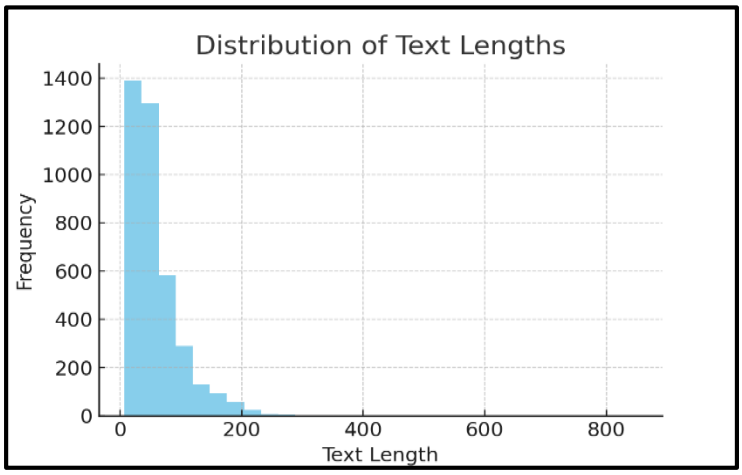


Fig. 3.3: Frequency Distribution of the Developed Corpora

Handling Code-Switching Challenge: It implements a word-weighted code-mixed method where code-switching is alternating languages in a singular conversation. The proposed function result for the sample opinion instance is illustrated in Fig. 3.4.

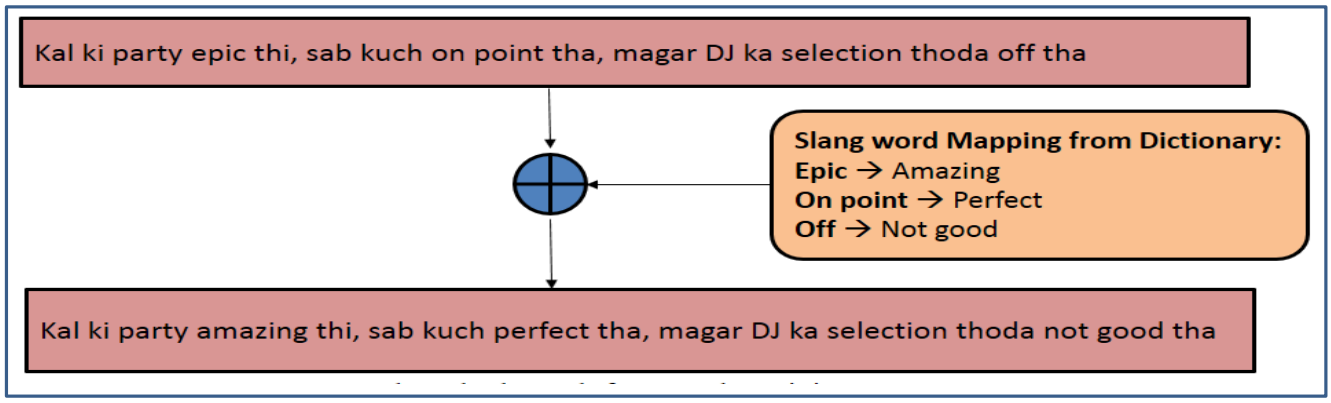

Kal ki party epic thi, sab kuch on point tha, magar DJ ka selection thoda off tha
Slang word Mapping from Dictionary:
Epic → Amazing
On point → Perfect
Off → Not good
Kal ki party amazing thi, sab kuch perfect tha, magar DJ ka selection thoda not good tha

Fig. 3.4: Result of Proposed Preprocessing Module for Sample Opinion

Handling Variations in Spelling and Code-Mixed Slang Words: Slang, stop words, and spelling variations are handled efficiently during pre-processing. Data was normalized for spelling variation, and slang was accurately interpreted through pre-processing. The proposed development of a corpus for code-mixed opinion mining handles limited resources, code-switching, spelling variations, and slang. For corpus development, opinions extraction through API calls and the Cohen Kappa measure for inter-annotation agreement (ITA) qualitative metric was implemented with an ITA of **94%**. Further, Slang Word Mapping and Replacement handles unstructured data in opinion text. We also developed a code-mixed stop word seed dictionary for effective stop word integration. The generated structured opinions are assessed through linguistic analysis using both quantitative and qualitative validation. The **qualitative validation demonstrates that the proposed corpus developed is linguistically variable, lexically complex, and suitable as input for opinion mining of unstructured data.**

CHAPTER 4

WEIGHTED CODE MIXED OPINION MINING FRAMEWORK (WCM-OMF)

4.1 OVERVIEW

Although there are studies related to opinion mining and opinion extraction, there is limited research to handle code switching issue in opinion corpus. There is a variation in actual meaning through accurate interpretation as compared to direct translation owing to lexical nuances of code-mixed opinions. To overcome this challenge, we propose the design and implementation of a weighted code-mixed opinion mining framework (WCM-OM) using an Adaptive Lemma-based weighted approach and machine learning for multi-class classification.

4.2. DATASET STATISTICS

This section describes the datasets utilized for the experimental evaluation of the proposed work, as explained in Table 4.1. Here, we employ two Benchmark datasets, Bohra et al. [43] and Joshi et al. [44], having Filtered data points 4575 and 3879, respectively. These benchmark datasets from the ACL data repository are validated through quantitative validation using the Cohens Kappa ITA score, making it suitable for statistical and qualitative performance validation for the proposed framework.

Table 4.1. Descriptive Statistics of the Benchmark Datasets used for Experimental Evaluation

Statistic	Bohra Dataset	Joshi Dataset
Total Data Points Collected	112718	4981
Filtered Data Points	4575	3879
Positive Class Data Points	1661 (Offensive Speech Opinions)	1358
Neutral Class Data Points	Not Available	1939
Negative Class Data Points	2914 (Normal Speech Opinions)	582
Inter Annotator Agreement (Cohen's Kappa)	0.982	0.64

4.3. PROPOSED OPINION MINING FRAMEWORK

This section presents the design and implementation of the proposed technique WCM-OM consisting of 6 significant modules: data acquisition, corpus extraction, data preprocessing for code-switched data, improved feature engineering module using Adaptive lemma-based vector space modeling, hybrid

opinion model development, and opinion detection through multi-class classification. The detailed components are described as follows:

a. Initially, we retrieve online data content extracted from heterogeneous web sources and unstructured data platforms from their respective programming APIs from benchmark datasets.

b. In the next phase, subtasks of NLP, including Stemming, Lemmatization, and irrelevant Noise Removal, are executed. Other subtasks may be required as per the application domain.

c. The step for detecting Language programmatically is crucial for efficient machine learning and deep learning model implementation.

d. The translation and transliteration involve conversion from 1 lingual and structural format to another, enabling efficient execution of the detection phase given the heterogeneity of opinions.

e. Feature extraction, transformation, and subsequent vectorization are executed by employing NLP in several diversified opinion classes.

f. Quantitative Performance evaluation is implemented to measure the effectiveness of training and test evaluation.

Fig. 4.1 explains the WCM-OM framework with the yellow highlighted cells representing the proposed solution contributions. Firstly, the pre-processed output from the prior module functions as input for WCM-OMF. Next, a language tagging module separates the opinion into parts of speech and tags for each sub-block. Subsequently, the ABIW lemma is implemented using Language Influence Weighting (LIW) methods, a proposed metric designed to capture the influence of various lexical contexts in the code-mixed input.

ABIW quantifies the extent to which the code-mixed input language contributes to the overall context. The linguistic feature is computed and combined with the statistical feature to obtain an EFV hybrid score. The code-mixed relevance score adjusts the TF-IDF score to get a new score. Weighting and normalization are applied to obtain the final Extended Feature Vector.

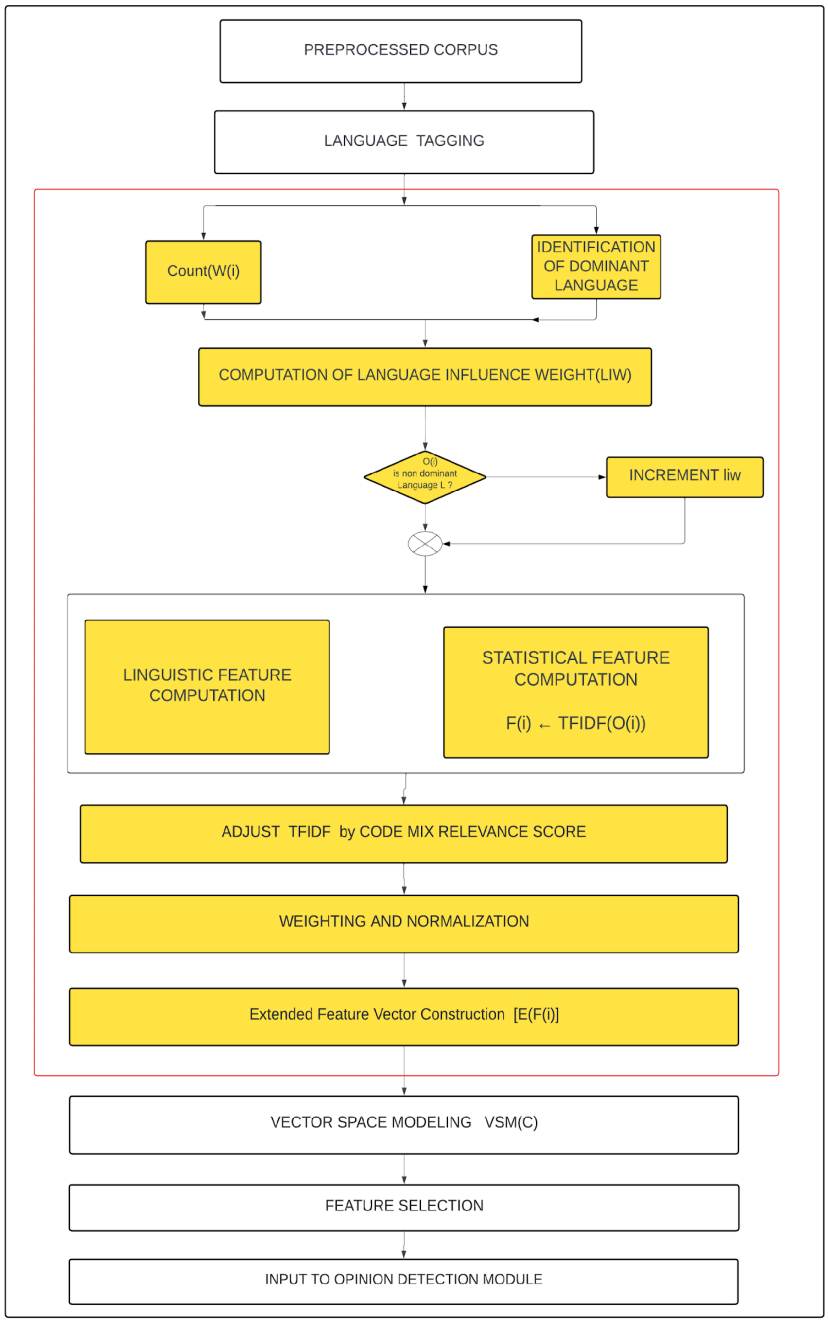


Fig. 4.1: Workflow of Proposed WCM-OMF Solution

4.4 WORKING MECHANISM OF WCM-OMF METHOD

The working mechanism of WCM-OMF, equations and mathematical formulation are described in this section. Equations 4.1, 4.2, and 4.3 explain the computation of term frequency and inverse document frequency.

$$TF(t,\ o)\ =\ \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ opinion\ o}{Total\ terms\ in\ \ opinion\ o} \tag{4.1}$$

$$IDF(t,\ O) = log\ \left(\frac{Overall\ opinion\ count}{Number\ of\ opinion\ with\ term\ t\ in\ it}\right) \quad (4.2)$$

$$TF - IDF = TF(t,o) * IDF(t,O) \quad (4.3)$$

Adaptive Bilingual Influence Weighting (ABIW), which is the language influence weighting method (LIW), is a metric designed to capture the influence of different languages present in the code-mixed text. It quantifies the extent to which each language contributes to the overall content. The proposed solution assigns weights based on the following:

- The Adaptive Bilingual Influence Weighting (ABIW)
- The number of code-switches in the initial sentence of the opinion.

For this purpose, the lemma ABIW is devised consisting of the following key components- EFV: Extended Feature Vector and TF-IDF: Term Frequency –Inverse Document Frequency. The weighting factor is computed for each opinion instance using a language identification function in combination with a weighting parameter. This result is added to the feature vector module, capturing the linguistic feature weights. The proposed lemma is explained in Equation 4.4, which dynamically adjusts the influence of each language based on the opinion.

$$ABIW = \ \sigma\left(\frac{1}{N} *\ \Sigma(i = 1\ to\ N) LangScore_i\right) \quad (4.4)$$

Where:

- W_{Lang} : weighting factor for each language in the bilingual code-mixed text
- $Model_m$: Trained model m
- N: the number of tokens in a segment,
- $LangScore_i$: score indicating the relevance of the language of the ith token,
- σ : normalization function.

Further, we implement vector space modeling on the extended feature vector for the feature selection process. Equation 4.5 computes the final feature vector for multiclass opinion classification and subsequent opinion detection.

$$CMRS(O(i)) = \ \sum_{i=0}^{n} LF(O(i)) + \ STF(O(i)) \quad (4.5)$$

Where:

- $CMRS$: Code Mixed Relevancy Score
- $O(i)$: i th Opinion where i Є [1,2,3…n] and n is the Total count of opinions
- STF –Statistical Features Set
- LF- Linguistic Feature Set

The proposed solution implements the computation of an Extended Feature Vector (EFV) instead of a Feature Vector (FV) in existing systems for feature engineering opinion detection. It consists of LIW computation, Feature Engineering, preparation of the dataset, prediction and performance parameters evaluation Further, the resultant output is then inputted for embedding and classification, enabling linguistic context integration that assists in distinguishing opinions having varied linguistic hedges.

4.5. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the experimental result outcome of the proposed WCM-OMF methodology for code mixed data input.

Table 4.2. Comparative analysis of Proposed WCM-OMF with existing systems on Benchmark Bohra Dataset

Methodology	Accuracy	Recall	F1-score
Support Vector Machine Model	70.7	31.3	0.429
Random Forest Model [43]	65.1	19.3	0.292
Sub-word LSTM model [43]	69.8	36.5	0.458
Hierarchical LSTM model + attention on phonemic sub-words [45]	66.6	45.1	0.487
WCM_OMF (Proposed weighted code-mixed approach)	**79**	**79.11**	**0.79102**

Table 4.2 results indicate that the proposed weighted code-mixed opinion mining solution significantly improves opinion classification accuracy using weighted strategy for code-switching implementing the k-fold cross-validation technique. Results show that the proposed WCM-OMF solution performs better than the existing system, achieving an accuracy of 79% for the Benchmark Bohra dataset. The Benchmark Dataset 1 is taken from Bohra, A. et al., Association for Computational Linguistics (ACL), sub-branch of ACM. Hindi-English code-mixed corpus is created using the tweets posted online in five years.

Table 4.3 results show that WCM-OMF performs significantly better than the existing opinion mining frameworks, with an overall accuracy achieved of 72.55% and an F1 score of 0.72.

Table 4.3. Comparative analysis of Proposed WCM-OMF with existing systems on Benchmark Joshi Dataset

Methodology	Authors	Accuracy	F1-score
BSVM (Unigram)	Wang and Manning [46]	59.15%	0.5335
NBSVM (Uni + Bigram)	Wang and Manning [46]	62.50%	0.5375
MNB (Unigram)	Wang and Manning [46]	66.75%	0.6143
MNB (Uni + Bigram)	Wang and Manning [46]	66.36%	0.6046
MNB (TFIDF)	Wang and Manning [46]	53.53%	0.4783
SVM (Unigram)	Pang and Lee [1]	57.60%	0.5232
SVM (bigram)	Pang and Lee [1]	52.96%	0.3773
Lexicon Lookup	Sharma et al. [47]	51.15%	0.252
LSTM Model 1	Joshi et al [44]	59.80%	0.511
LSTM Model 2	Joshi et al [44]	59.70%	0.658
WCM_OMF (Proposed weighted code-mixed approach)	**Proposed**	**72.55%**	**0.7208**

The Opinion selected consists of specific hashtags and keywords from politics, public protests, etc., which have a good propensity for the presence of online unstructured data. Approximately 1,12,718 tweets from Twitter in JSON format are considered during initial extraction, which consists of information such as timestamp, URL, text, user, re-tweets, replies, full name, id, and likes. The proposed solution WCM-OMF results are designed and implemented to assign weights based on the ABIW lemma and the number of code switches in intra-sentential opinion. The proposed WCM-OMF model achieves an F1 score of **0.79** on benchmark Bohra dataset, achieving a **6%** significant improvement in opinion mining framework over existing state-of-the-art-methods.

CHAPTER 5

FENNEC FOX OPTIMIZATION OPINION MINING FRAMEWORK (FFO-OMF) USING HYBRID APPROACH

5.1 OVERVIEW

The lexical issue of opinion mining described in previous chapters affects NLP operations, including opinion analysis, since mixed language construct opinions are challenging to capture. This chapter presents the proposed Fennec Fox Optimization Opinion Mining Framework (FFO-OMF) using a hybrid approach of a bioinspired metaheuristic-based technique integrated into an opinion mining system for multiclass classification. The NFL (no free lunch) theorem [48] [49] emphasizes nature-inspired algorithms. The proposed solution adapts the Fennec Fox Optimization (FFO) algorithm[50] to implement the opinion mining feature-engineering module. The proposed framework implements local search using the simulated fox's efficient search and escape technique. The FFO's optimization methodology [86] replicates complex search environments. Deterministic strategies [51] are inappropriate for real-world optimization problems due to high dimensionality, stochastic objective functions, complex search spaces, and multiple local optimal performance [52]. Due to unpredictability, balancing exploration and exploitation operations is challenging to implement—the exploration phase searches to avoid early convergence to inadequate solutions.

5.2 WORKING OF PROPOSED FFO- OMF METHODOLOGY

The FFO-OMF steps are explained in detail in this section.

Step 1 - Initialization: Generate the initial feature set $\boldsymbol{F(X_i)}$ with each $\boldsymbol{f_i}$ being a potential attribute

$$F = \begin{pmatrix} F(X_1) \\ \vdots \\ F(X_N) \end{pmatrix} \tag{5.1}$$

Population Matrix Representation

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,m} \end{pmatrix} \tag{5.2}$$

Step 2- Evaluation: Use a fitness function F_{fs} to determine the relevance of each feature in the opinion feature selection phase.

$$F_{\text{fs}} = [F_{\text{fs},1}, F_{\text{fs},2}, \dots, F_{\text{fs},N}]_{N\times 1} \tag{5.3}$$

Step 3 - Searching: Iteratively refine feature sets with Fitness$(\boldsymbol{f_i})$ to find optimal combinations for accuracy.

$$x_{P1i,j} = x_{i,j} + (2 \cdot r - 1) \cdot R_{\text{fs},j} \tag{5.4}$$

(where $i = 1,2,\dots,N$ and $j = 1,2,\dots,m$)

Neighbourhood Radius Defined for Features Representation

$$R_{\mathrm{fs},j} = \alpha \times \left(1 - \frac{t}{T}\right) \tag{5.5}$$

Step 4 - Adaptation: Adjust search strategies based on feature combination successes, mimicking fennec fox learning for better feature refinement with Conditional Feature Update

$$X_{\mathrm{fs}} = \begin{cases} X_{\mathrm{fs}}^{P1} & \text{if } F_{\mathrm{fs}}^{P1} > F_{\mathrm{fs}} \\ X_{\mathrm{fs}} & \text{otherwise} \end{cases} \tag{5.6}$$

Step 5 - Communication: Share successful feature sets across iterations to guide FFO toward the most effective feature set for opinion mining. Final Feature Update is given as:

$$X_{\mathrm{fs}} = \begin{cases} X_{\mathrm{fs}}^{P2} & \text{if } F_{\mathrm{fs}}^{P2} > F_{\mathrm{fs}} \\ X_{\mathrm{fs}} & \text{otherwise} \end{cases} \tag{5.7}$$

Table 5.1: Mathematical Notation and Description in FFO-OMF

Notation	Description
FF	Fennec Fox
$\boldsymbol{x_{i,j}}$	The j^{th}dimension of the i^{th} FF
$\boldsymbol{lb_j}$	Lower bound of the j^{th}dimension
$\boldsymbol{ub_j}$	Upper bound of the j^{th}dimension
$\boldsymbol{r}$	Random number generated in the range from [0, 1]
$\boldsymbol{X}$	The population matrix consists of N number of FF.
$\boldsymbol{m}$	The total number of decision variables
$\boldsymbol{N}$	The total number of Fennec Foxes in the population
$\boldsymbol{F(X_i)}$	The objective function value for the i^{th} FF
$\boldsymbol{R_{fs,j}}$	The neighborhood radius for the j^{th} dimension of the i^{th} FF
$\boldsymbol{t}$	The current iteration counter
$\boldsymbol{T}$	The total number of iterations in the algorithm
$\boldsymbol{\alpha}$	A constant used in the exploitation phase (generally set to 0.2)
$\boldsymbol{X_{\{P1,i\}}}$	The newly suggested position of the i^{th} FF based on the first phase
$\boldsymbol{F_{\{P1,i\}}}$	The objective function value for the i^{th} FF in the first phase
$\boldsymbol{X_{rand,i}}$	The target position intended for the escape of the i^{th} FF

$F_{rand,i}$	The objective function value for the randomly chosen position in the exploration phase
I	A random number selected from the set {1, 2}
$X_{P2,i}$	The newly suggested position of the i^{th} FF based on the second phase
$F_{P2,i}$	The objective function value for the i^{th} FF in the second phase

Table 5.1 shows the mathematical notations and the corresponding description of the attributes of the FFO algorithm.

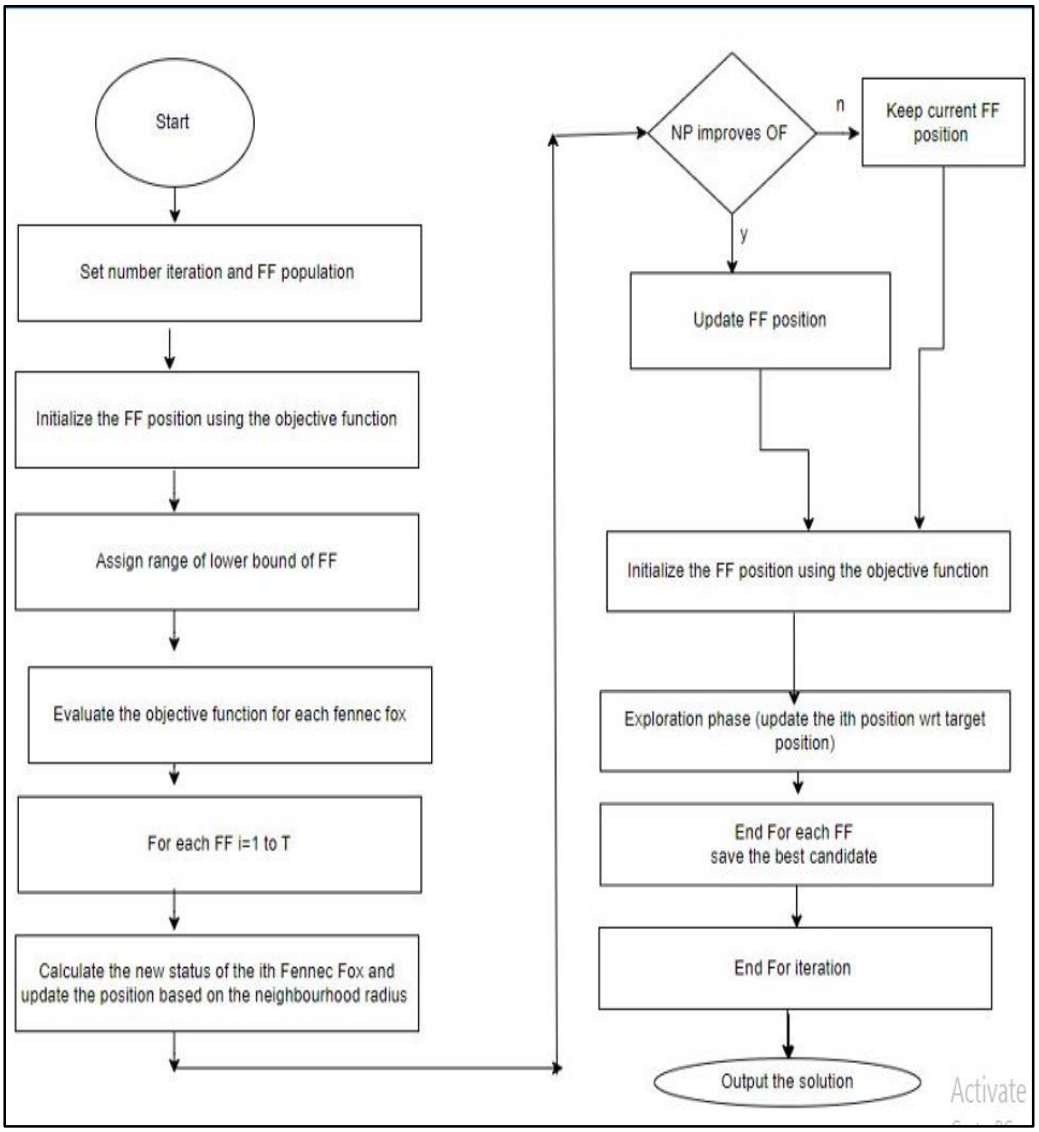

Start
Set number iteration and FF population
Initialize the FF position using the objective function
Assign range of lower bound of FF
Evaluate the objective function for each fennec fox
For each FF i=1 to T
Calculate the new status of the ith Fennec Fox and update the position based on the neighbourhood radius
NP improves OF
n
Keep current FF position
y
Update FF position
Initialize the FF position using the objective function
Exploration phase (update the ith position wrt target position)
End For each FF
save the best candidate
End For iteration
Output the solution
Activate

Fig. 5.1: Workflow of the Proposed FFO-OMF Solution

This is the first attempt to hybridize FFO with an opinion detection module to the best of the author's knowledge. Fig. 5.1 illustrates the flow graph for FFO-OMF. Here, NP: New position of the Fennec

Fox; T: Total Foxes; OF: Opinion Framework. NP updates the FF position if it improves the candidate solution; it retains the original position in the neighborhood radius. Post this; the best candidate solution is updated. The updated solution of candidates generated from the module of Fennec Fox is provided as input to the opinion classification module for further detailed processing in opinion detection. This ensures that the parameters from the Fennec Fox algorithm module enhance the accuracy and precision of opinion classification during the opinion mining phase of proposed FFO-OMF.

5.3 EXPERIMENTAL RESULTS AND DISCUSSION

FFO_RF and FFO_XGB are proposed hybrid models employing FFO with Random Forest and Extreme Gradient Boosting, respectively. Fig. 5.2 depicts the comparative analysis of the Proposed FFO-OMF solution on the Bohra Benchmark [77] dataset.

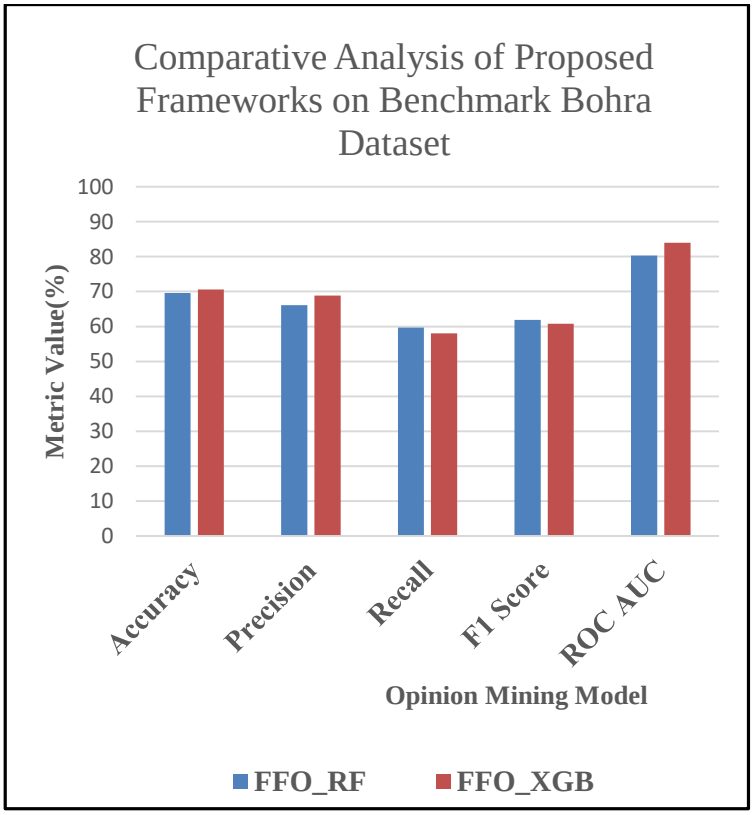


Fig. 5.2 Performance Analysis of the Proposed FFO-OM Framework on Benchmark Bohra Dataset

Results show that the proposed FFO-OMF solution achieves significant results for majority of the evaluation metrics in the context of unstructured data.

CHAPTER 6

ADAPTIVE FENNEC FOX OPTIMIZATION OPINION MINING FRAMEWORK (AFFO-OMF)

6.1 BACKGROUND

The existing systems, including mathematical programming, gradient-based methods, and metaheuristic algorithms, exhibit significant limitations, resulting in an unbalanced search mechanism. Stochastic techniques [53] exhibit self-organization, simplicity, and distribution properties. Recent metaheuristic algorithms include Social Spider optimization, Cuckoo-based metaheuristic techniques, amongst others [54][55]. These algorithms generally perform poorly when applied across diverse optimization problems. The proposed solution AFFO-OMF leverages the Levy flight concept within the context of the code mixed corpus provided.

a. Random Walk

A Random Walk (RW) is a stochastic process [56] having arbitrary steps, defined as:

$$S_N = S_{N-1} + X_N \tag{6.1}$$

The next stage S_N depends on the current state S_{N-1} and next state transition. RW search space exploration strategy; however, a simple or "blind" RW is inefficient .

b. Distribution of Levy Flight

The levy distribution is a heavy-tailed probability distribution function that may result in substantial shifts throughout the search region, facilitating a global solution. Levy-Flight's [57], deliberate leaps assist the FFO algorithm in avoiding the premature convergence state. The Levy distribution is a generalization of the Gaussian distribution function formulation as below:

$$L(s) \sim |s|^{-1-\beta} \tag{6.2}$$

$where\ 0 < \beta < 2$ denotes the index, and s represents the step length.

$$s = \frac{p}{|q|^{\frac{1}{\beta}}} \tag{6.3}$$

6.2 PROPOSED METHODOLOGY

The Adaptive Fennec Fox Optimization Opinion Mining Framework extends the proposed FFO-OMF framework, integrating bio-inspired optimization for the feature selection module. The proposed AFFO-OMF framework efficiently balances the exploration phase and the exploitation phase. Further, hybridization of the feature selection of Fennec Fox and embedding technique is implemented, enabling AFFO-OMF to perform well in unstructured code mixed opinion detection. The problem of local

optima that often hinders traditional optimization methods is overcome in the proposed method by introducing the non-Gaussian Stochastic Levy function. The key strength of the Levy flight mechanism lies in its ability to introduce infrequent leaps in the search space for escaping the local entrapment. Furthermore, this refinement in AFFO-OMF leads to a more effective opinion mining framework capable of delivering superior performance across diverse linguistic contexts and data variability. Fig. 6.1 describes the workflow of the proposed AFFO-OMF solution. Here, the yellow highlighted blocks represent contribution.

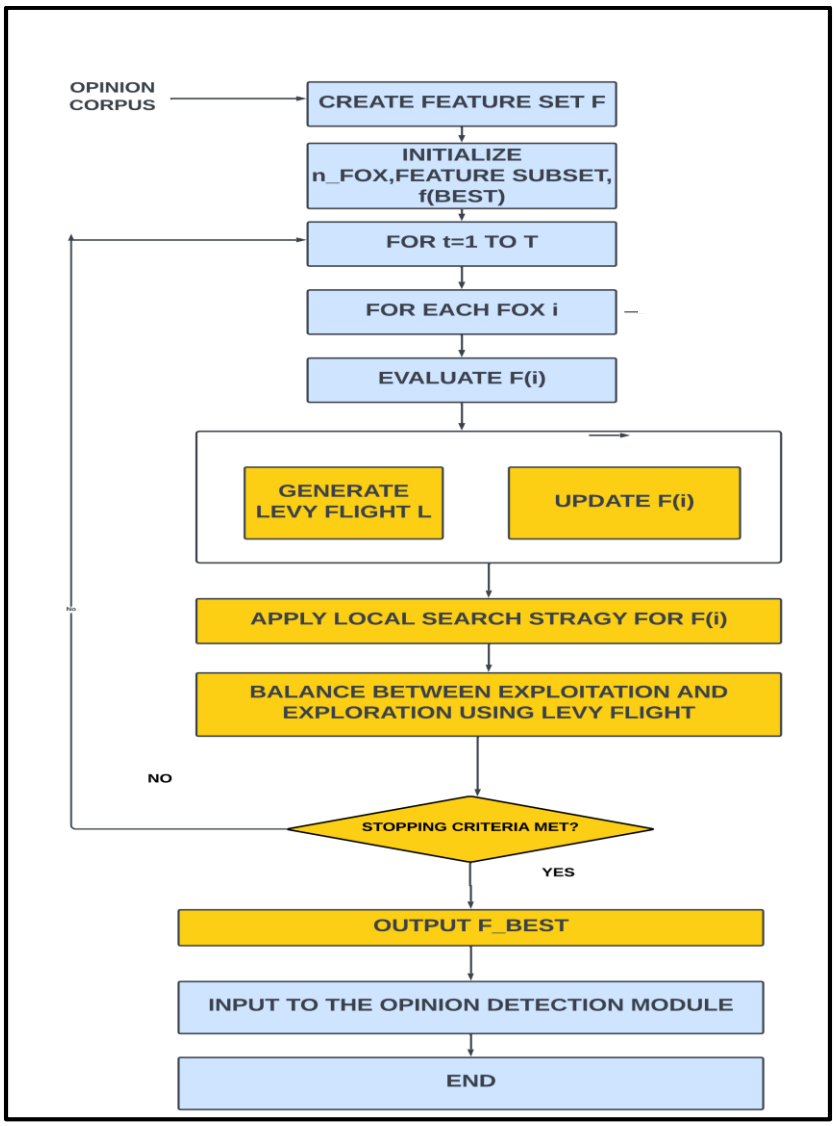


Fig. 6.1: Proposed Adaptive Fennec Fox Based Opinion Mining Framework (AFFO-OMF)

6.3 EXPERIMENTAL RESULTS

The proposed technique was implemented using Python implementation on an Intel i5 processor and 8 GB of RAM. Results show that the adaptive nature of AFFO-OMF, driven by Levy flights, is effective, particularly for linguistically challenging code mixed environments. The combination of the fennec fox-inspired strategies with Levy flight behaviour enhances the algorithm's global exploration

capabilities and potentially improve its convergence rate and solution quality. The proposed technique was evaluated utilizing standard evaluation parameters- accuracy, precision, recall, and F1-score.

For Benchmark Dataset 1: The benchmark corpora for the proposed model evaluation were taken from Bohra et al., Association for Computational Linguistics (ACL), ACM sub-branch

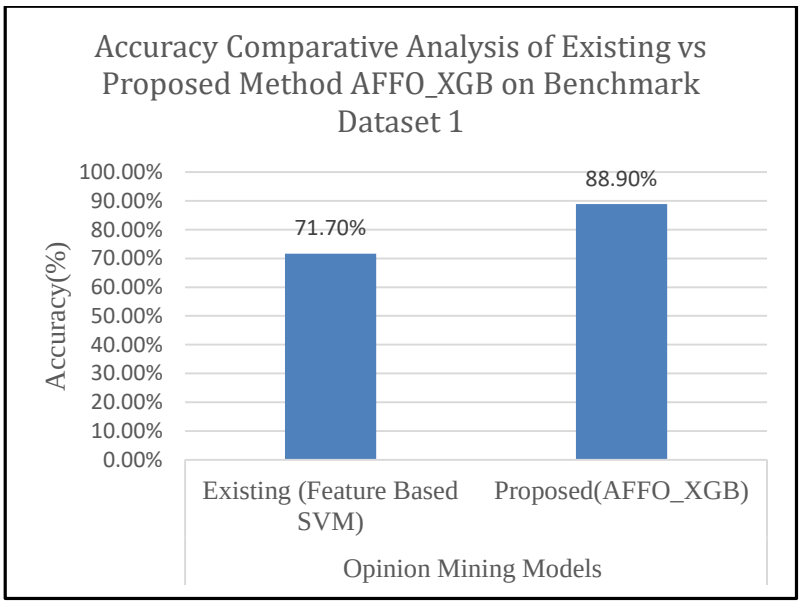


Fig. 6.2: Experimental Result on Benchmark Dataset 1 for Proposed AFFO-OMF

Results show that the proposed model AFFO_XGB achieves an efficient accuracy of **88.9%.** AFFO-OMF is implemented for both Benchmark Dataset 1(Bohra et al.) [77]. Fig. 6.2 demonstrates proposed framework AFFO_XGB, performing significantly better than the existing Feature Based SVM model.

For Benchmark Dataset 2: Fig. 6.3 shows that the proposed methodology AFFO_XGB performs significantly better than existing Multinomial Naïve Bayes with unigram on Benchmark corpora, Joshi et al., ACL [44]. AFFO_XGB achieves an improved accuracy of **72.70%** for Benchmark Dataset 2.

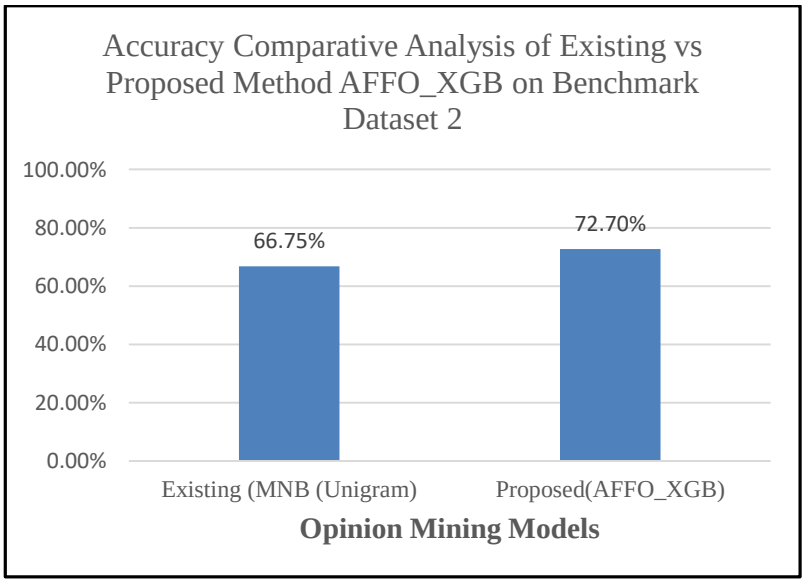


Fig. 6.3: Experimental Result on Benchmark Dataset 2 for Proposed AFFO-OMF

Table 6.1 demonstrates that the proposed approach, AFFO+XGB, achieved a precision rate of **89.32%** and recall of **88.9%**, outperforming existing models such as Hierarchical LSTM, MOH + MBERT (2022) and MoH+MuRIL (2022) on Benchmark Dataset 1.

Table 6.1: Comparative Analysis of Proposed AFFO model with existing state of the art methods

Model Name	Performance Parameter		
	Precision	**Recall**	**F1**
Synthetic + Gold (Samanta et al., 2019) [58]	59%	63%	53%
Hierarchical LSTM (Santosh & Aravind, 2019) [45]	–	45%	49%
CNN-1D (Kamble & Joshi, 2018) [59]	83%	78%	80%
MoH + M-Bert (Arushi S & Kabra, 2022) [60]	83%	80%	81%
MoH + MuRIL (Arushi S & Kabra, 2022) [60]	86%	84%	85%
Proposed model (AFFO+ XGB)	**89.32%**	**88.97%**	**89.02%**

6.4 FINDINGS AND DISCUSSION

Results show the comparison of the proposed solution with other existing techniques, including Synthetic + Gold (Samanta et al., 2019), for code-switching data without grammar conversion. Another existing opinion detection system implements character embeddings through 1-D sub-word level representations. Hierarchical LSTM (Santosh & Aravind, 2019) utilizes attention-based mechanism and phonemic based sub-words. Other existing research [58,59] has experimented with domain-specific word embeddings on deep learning techniques. Here, two models are further proposed and presented as follows:

a. FFO-OMF (Fennec Fox Optimized Opinion Mining Framework) – Implemented in Chapter 5 with FFO-RF: FFO integrated with Random Forest Algorithm and FFO_XGBOOST: FFO integrated with Extreme Gradient Boosting
b. AFFO-OMF (Adaptive Fennec Fox Optimized Opinion Mining Framework) with AFFO_XGB: AFFO integrated with Extreme Gradient Boosting

Fig. 6.4 illustrates the proposed opinion detection system (AFFO + XGB) performance analysis implemented considering various performance metrics. Out of the proposed frameworks, AFFO_XGB gives the best performance, achieving an accuracy of **88.9%.**

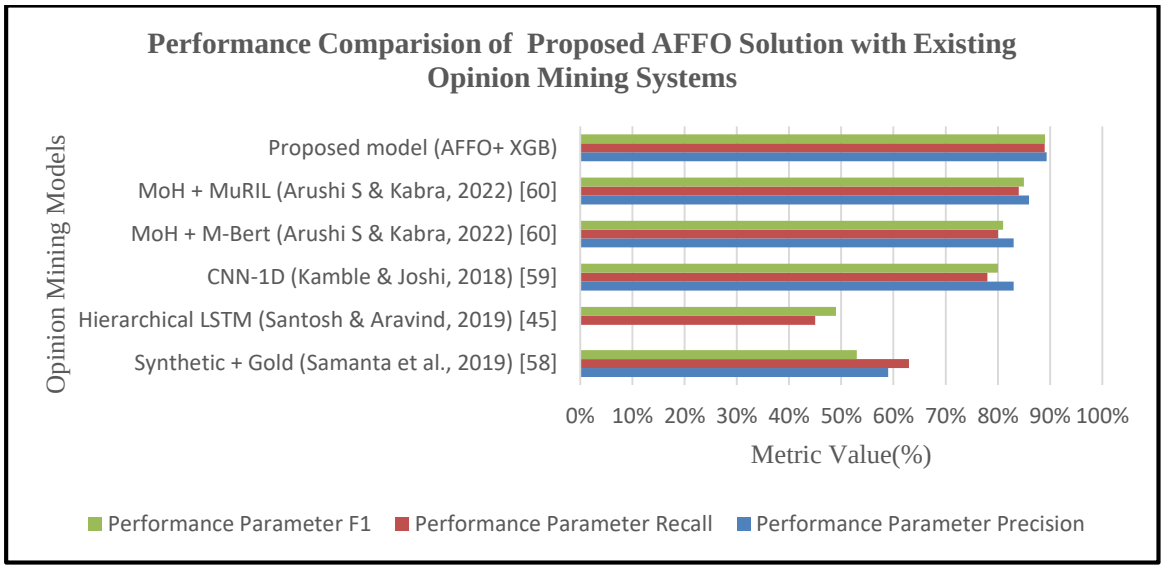

Performance Comparision of Proposed AFFO Solution with Existing Opinion Mining Systems
Opinion Mining Models
Proposed model (AFFO+ XGB)
MoH + MuRIL (Arushi S & Kabra, 2022) [60]
MoH + M-Bert (Arushi S & Kabra, 2022) [60]
CNN-1D (Kamble & Joshi, 2018) [59]
Hierarchical LSTM (Santosh & Aravind, 2019) [45]
Synthetic + Gold (Samanta et al., 2019) [58]
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Metric Value(%)
Performance Parameter F1
Performance Parameter Recall
Performance Parameter Precision

Fig. 6.4: Comparative Analysis of Proposed AFFO Solution with Existing Opinion Mining Systems

Existing systems utilize F1, Recall, and Precision, while researchers [45] have demonstrated their model through F1 and Recall performance metrics. This proposed method's primary contributions are:

a. AFFO-OMF, a metaheuristic approach for maximizing feature selection and model parameters in opinion classification for code-mixed text improves FFO to balance exploration and exploitation.

b. The results show that AFFO-OMF performs significantly better than existing state-of-the-art (SOTA) systems on code mixed benchmark datasets and for majority of the critical performance metrics. AFFO-OMF is the best-performing model among the proposed models, achieving an accuracy of 88.9% and a precision of 89.32%.

CHAPTER 7

ENSEMBLE CODE MIXED OPINION MINING FRAMEWORK (ENCOMF)

7.1 INTRODUCTION

The contributions of this chapter are:

- Detailed benchmarking of proposed RHIEN corpora using standardized machine learning algorithms.
- ENCOMF –an ensemble opinion mining framework for opinion detection is proposed and compared with related existing approaches
- Further in-depth validation of proposed model results is implemented using various performance metrics utilizing a modified Bayesian Optimization approach.

7.2 PROPOSED METHODOLOGY

The proposed approach implements a hybrid framework of natural language processing, machine learning, and transformer approaches.

PHASE I - DATA ACQUISITION: Sample opinions were extracted with annotation H representing hate and NH representing non-hate.

PHASE II - FEATURE ENGINEERING: In the proposed method, we construct Class C= "H" or "NH." Here, H = Hate-based opinion and NH= Non-hate-based opinion. A corpus entry e = <t, c> is a tweet and its class label. An ordered set, r, includes the complete opinion and its annotation. D = $\{O_1, O_2, O_3, \ldots . O_n\}$, where |n| = the total number of opinions. The emojis in the opinion corpus are removed. Any null opinions, duplicate opinions found in the corpus, URLs are eliminated, and opinions are then converted to lowercase for normalization. On source corpus SC, the stemming and lemmatization processes are implemented. Consequently, we implement the stop word removal process for every $O_i \subseteq$ D. Next, the Term Frequency Inverse Document Frequency (TFIDF) is implemented in feature engineering step and the code mixed opinions are segmented into mixed tokens in tokenization process. In Lemmatization step, the opinion is stripped and converted to the opinion root format.. The TF scores would be computed based on their corresponding frequencies. Inverse Document Frequency (IDF) is computed as a weighted matrix. Thereafter, we compute the matrix and score representation for the source corpus for opinion classification . The classification hyperparameter bounds selected for SVM are svm_param_bounds ← {'C': (0.1, 10), 'gamma': (0.1, 1)}

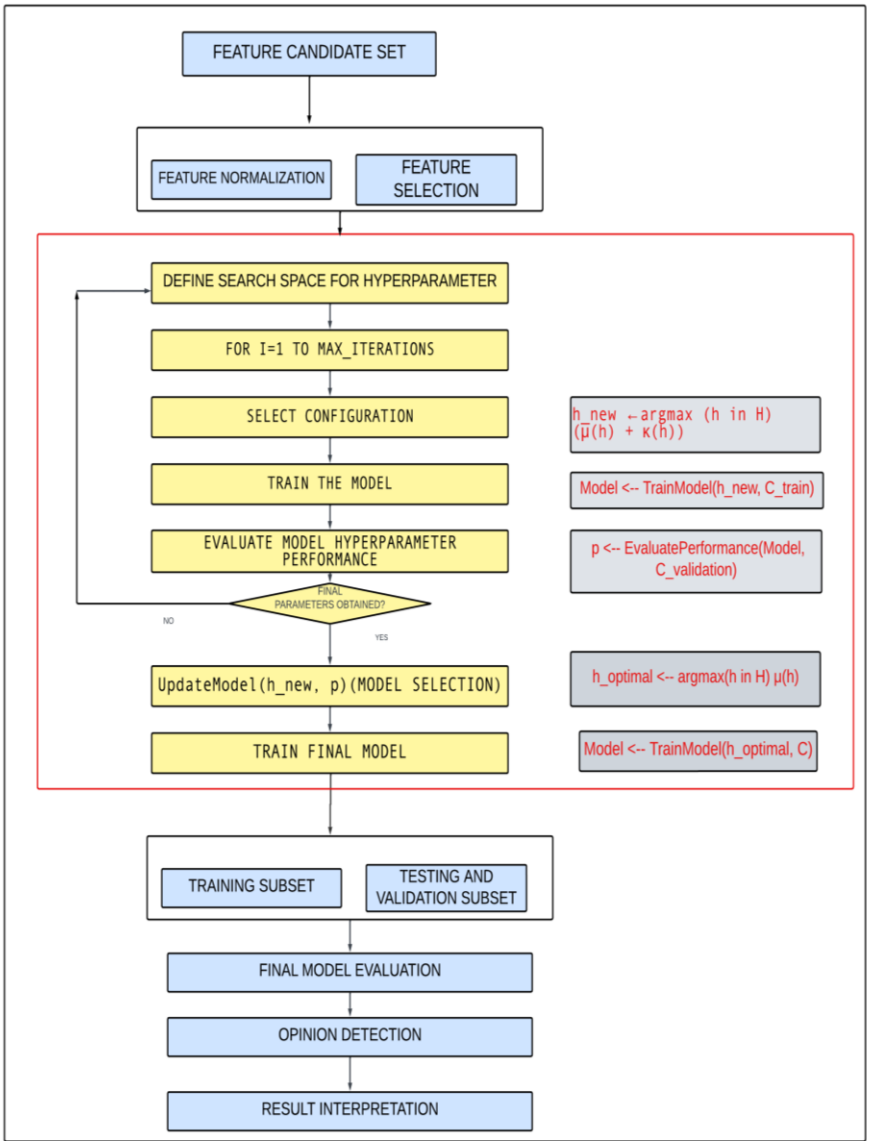


Fig. 7.1: Proposed ENCOMF Solution Workflow

Fig 7.1 illustrates the steps of the proposed ENCOMF workflow described as follows:

a. Define the search space for hyperparameters: H

b. Initialize the probabilistic model for search space based on the initial observations.

c. Repeat for i = 1 to max_iterations:

   i. Select the subsequent hyper parameter configuration to evaluate

   ii. Select the configuration with the highest acquisition function value

   iii. Train the opinion analysis model using $hnew$ and C_{train}(Training corpus)

   iv. Employing a validation set, assess the model's performance p

   v. Updation of the Predictive model Pi with the new observation

   vi. Select the configuration with the highest UCB performance

Here, The UCB (Upper Confidence Bound) equation is defined as:

$$\kappa\,(h) = \beta * \sqrt{\ln(n)\,/n_h} \qquad (7.1)$$

Where:

- κ(h): The UCB value for a hyperparameter configuration h.
- β: A hyperparameter that governs the balance of exploration and exploitation.
- n: Total number of iterations.
- n_h: Number of times the hyperparameter configuration h has been sampled.
- $sqrt\left(\frac{\ln(n)}{n_h}\right)$: The confidence interval around the estimated performance of h based on the number of samples.

d. Train the opinion mining model with the optimal hyper parameters computed

e. Train and evaluate the final model

Phase III implements model evaluation and corresponding opinion detection results explained in next section.

7.3 EXPERIMENTAL RESULTS

In this section, we have evaluated the experimental design proposed for ENCOMF. The proposed technique was implemented using Python programming on an Intel i5 processor and 8 GB of RAM. The recommended optimization technique utilizing accuracy, precision, recall, and F1-score quantitative metrics was thoroughly evaluated.

$$O_{ACC} = \frac{\text{TPO} + \text{TNO}}{\text{TPO} + \text{TNO} + \text{FPO} + \text{FNO}} \tag{7.2}$$

Where: TPO – True positive Opinion; TNO – True Negative Opinion

FPO – False positive Opinion; FNO – False Negative Opinion

Table 7.1: Experimental Results of Proposed ENCOMF Using Hybridization of Transformer Based Learning and Optimization

Machine Learning Techniques	**Experimental Results(%)**					
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Kappa**	**AUC**
Multinomial Naive Bayes	83.5	81.2	91.6	86	66	82.4
Decision Tree	86.6	88.8	86.9	87.8	72.9	86.6
KNN	59.5	58	99.3	73.2	99	54.5
Proposed Ensemble + Hard Voting	88.7	87.8	92.5	90.1	76.9	88.2
Proposed Ensemble + Soft Voting	88.3	86.2	94	89.9	76	87.6
Proposed Ensemble + Transformer (BERT based learning)	85.71	1	75	85.714	72	87.5

Table 7.1 describes the experimental results of the existing Machine Learning Algorithms, the proposed ensemble approach, and the Transformer Learning approach. Table 7.2 shows the comparative analysis of Proposed Method ENCOMF variants implemented for Benchmark dataset 1. Here, the proposed ENCOMF variant is implemented in hybridization with Extreme Gradient Boosting (XGBOOST), Linear Kernel method (SVM_Linear), and Radial Basis Function kernel (SVM_RBF) method. Results indicate that the ENCOMF+SVM_Linear proposed approach gives the best result, achieving an accuracy level of **94.34%.**

Table 7.2 Proposed ENCOMF variant results on Benchmark Bohra dataset.

Methodology (ENCOMF variants)	**Performance Metrics(%)**				
	Accuracy	**Precision**	**Recall**	**Fl**	**ROC AUC**
ENCOMF + XGBOOST	81.91	82.92	81.92	82.04	94.43
ENCOMF + SVM_Linear	94.34	94.40	94.331	94.33	98.74
ENCOMF + SVM_RBF	90.22	90.31	90.214	90.22	97.66

The optimization values for every iteration I1 to I5 of the proposed ensemble model are depicted in Table 7.3. The proposed ENCOMF model generates iteration values for each learning target for target, learning rate, and number of estimators N_{Est}.

Table 7.3 Iteration Values for Optimization in Proposed ENCOMF

SrNo	**Iteration Results**			
	Iteration	**Target**	**Learning**	N_{Est}
1	I1 |	-0.8868	0.3808	477.8
2	**I2**	**-0.88**	**0.7347**	**319.4**
3	I3	-0.8953	0.1645	120.2
4	I4	-0.9004	0.0675	439.8
5	I5	-0.8896	0.6051	368.6

Experimental Results of the proposed ENCOMF solution on the developed RHIEN corpus are presented in Fig. 7.2.

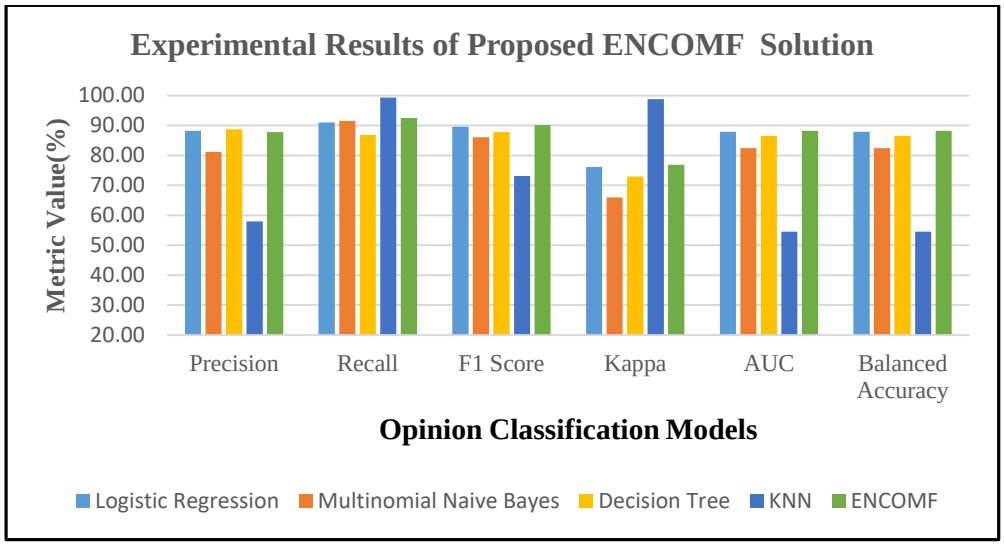


Fig. 7.2: Proposed ENCOMF Comparison with Existing Machine Learning Techniques on Developed Corpus

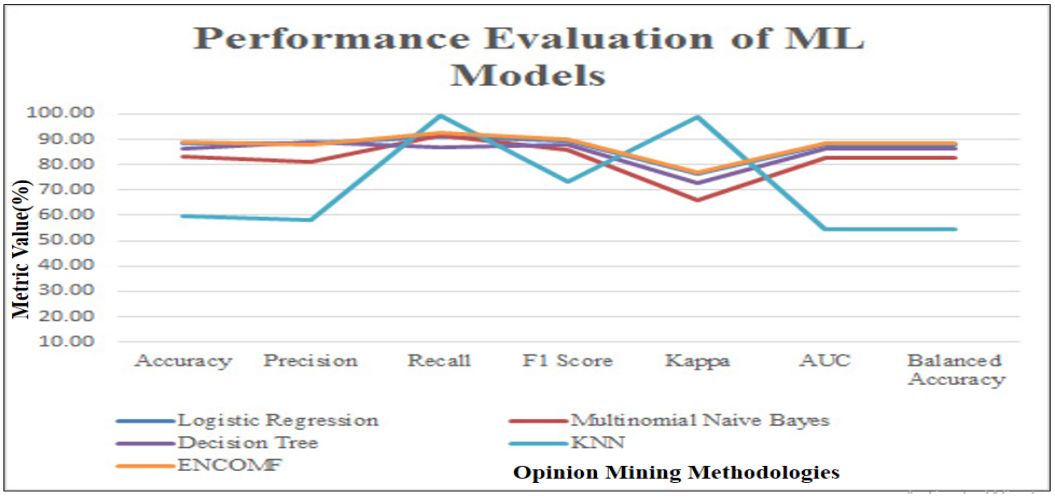


Fig. 7.3: Performance Comparison of Proposed ENCOMF Solution with standard Machine Learning Techniques

7.4 FINDINGS AND DISCUSSION

As observed in Fig. 7.3, the proposed approach performs significantly better than the existing standard Machine Learning algorithms for Opinion Mining across the evaluation metrics.ENCOMF- ensemble opinion mining framework in the context of code-mixed opinions and unstructured format implements benchmarking of the developed corpus and hyper-parameter-based model selection Results show that the proposed ENCOMF performs better than existing standardized machine learning techniques, achieving an accuracy of **88.7%** and recall of **92.5%,** leading to enhanced decision-making.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 CONCLUSION

The following conclusions were drawn from this research work:

a. Through the extensive literature review, we observed the need to develop a corpus for code-mixed data due to low resource availability. To address this research gap, we initiated our research work with the design and development of **Romanized Hindi English Opinion Corpus (RHIEN)**. We further performed annotation and statistically validated the developed corpus using quantitative metrics and linguistic metric analysis.

b. In our following implementation, we propose the **Weighted Code-Mixed Opinion Mining Framework (WCM-OMF)** to overcome the challenge of the code-switching problem in linguistic corpora. The performance of WCM-OMF improves by applying the proposed weighted approach on unstructured corpora through the implementation of the adaptive bilingual lemma proposed.

c. Further, to analyze the effectiveness of bio-inspired metaheuristic algorithms on code-mixed text, we developed the **FFO-OMF (Fennec Fox Optimization-based Opinion Mining Framework)**. To the best of our knowledge, this is the first attempt to integrate hybridized FFO for code-mixed analysis. It is observed that FFO performs well for feature engineering.

d. The proposed framework, **Adaptive Fennec Fox Optimization Opinion Mining Framework (AFFO-OMF),** overcomes the issue of local optima. We implemented the levy flight mechanism for the feature selection process that provides a balance between the exploration and the exploitation phase of the proposed framework. AFFO achieved an **F1 score of 88.91%**, performing significantly better than the state-of-the-art techniques.

e. We tested the proposed solutions **WCM** and **AFFO** over two benchmark datasets. It is observed that all proposed solutions perform better than existing opinion detection systems on both benchmark datasets and across the majority of the performance metrics. The improved results are achieved using individual classifiers without taking the support of ensemble techniques like multiple classifiers voting, blending, and stacking, which enhances performance.

f. Further, we develop an end-to-end framework using an ensemble of machine learning and natural language processing on our developed corpus. The proposed framework, **Ensemble Code Mixed Opinion Mining Framework (ENCOMF)** overcomes the problem of the lack of benchmarking on several machine learning algorithms and for all the standard metrics. The

benchmarking results show that the proposed framework achieved an **accuracy of 89.02%,** enabling business intelligence for decision-making.

8.2 CONTRIBUTION SUMMARY

Table 8.1 presents the research contributions summarized in a tabular format.

Table 8.1: Summary of the Research Contributions

SrNo	Proposed Solutions	Purpose
1	Corpus Development of RHIEN— Romanized Hindi + English Code-Mixed Corpus	To solve the problem of Data Sparsity and Low Resource Availability (LRA).
2	Design and Implementation of Weighted Code-Mixed Opinion Mining Framework (WCM-OMF)	To Handle Feature Engineering problem.
3	Design and Implementation of Fennec Fox Optimized Opinion Mining Framework (FFO-OMF)	
4	Design and Implementation of Adaptive Fennec Fox Optimized Opinion Mining Framework (AFFO-OMF)	
5	Design and Implementation of Ensemble Code Mixed Opinion Mining Framework (ENCOMF)	For Opinion Detection (Hybridization of NLP and Machine Learning) and Benchmarking.
6	Development of Lemma - Adaptive Bilingual Influence Weighting (ABIW)	To Handle Code Switching problem.
7	Development of functions for slang word integration, stop word generation and integration, and preservation of special characters for unstructured opinions.	To Handle Unstructured Data problem.

8.3 FUTURE WORK

The future enhancements in the submitted work are described as follows:

a. **Transfer learning** between high-resource and low-resource languages can be explored.
b. **Multimodal Opinion Mining**
c. **Comprehensive Opinion Datasets** to improve model resilience.

LIST OF PUBLICATIONS

1. Ruchi Sharma, Pravin Shrinath, "Improved Opinion Mining for Unstructured Data Using Machine Learning enabling Business Intelligence," JAIT- V14N4, Journal of Advances in Information Technology, Engineering Technology Publications, August 2023, (**Web of Science and Scopus**), Scopus Cite score- 3.1
2. Ruchi Sharma, Pravin Shrinath, "Ensemble of Weighted Code Mixed Feature Engineering and Machine Learning Based Multiclass Classification for Enhanced Opinion Mining for Unstructured Data" , IJACSA(**Web of Science and Scopus**) Accepted and Publication in press
3. Ruchi Sharma, Pravin Shrinath, "Leveraging Ensemble Opinion Mining Framework using Machine learning, "International Journal on Recent and Innovation Trends in Computing and Communication, 2023, Auricle Publishing House, (**Scopus** Indexed) – Accepted and Publication in press
4. Ruchi Sharma, Pravin Shrinath, "Bibliometric and Meta-analysis framework for code mixed opinions in unstructured format "revised paper submitted in JCCE, Journal of Computational and Cognitive Engineering, 2024 (**Scopus** Indexed – **Q1**) (under 2nd review)
5. Ruchi Sharma, Dr. Pravin Srinath, "Business Intelligence using Machine Learning and Data Mining- An Analysis," International Conference ICECA, IEEE, 2018. IEEE Explore (**Scopus** Indexed) -21 citations
6. Ruchi Sharma, Dr. Pravin Srinath, "Enterprise project management digitization using Machine learning," Scopus Indexed IEEE conference, IEEE Explore, 2018 (**Scopus** Indexed) - 27 citations
7. Ruchi Sharma, Pravin Shrinath, "Design and Implementation of Adaptive Metaheuristic Optimization based Opinion Mining framework using novel corpus," Copyright filing and publication in process.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[2] Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches, and applications." Knowledge-based Systems 89,2015: pp 14-46

[3] C. Barnali, "Code-Switching and Mixing in Communication- A Study on Language Contact in Indian Media" in The Future of Ethics Education and Research, Scientia Research Institute, pp. 110-123, 2017

[4] "Tapping the power of unstructured data," MIT Sloan, Sep. 2023. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data. [Accessed: Dec, 2023

[5] R. Sharma and P. Srinath, "Business Intelligence using Machine Learning and Data Mining techniques - An analysis," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1473

[6] B. Liu, Sentiment Analysis and Opinion Mining. Cham, Switzerland: Springer Nature, 2022.

[7] Sharma, Ruchi, and Pravin Shrinath. "Improved Opinion Mining for Unstructured Data Using Machine Learning Enabling Business Intelligence." Journal of Advances in Information Technology 14, no. 4 (2023): 821-829.

[8] Verma, Sanjeev. "Sentiment Analysis of Public Services for Smart Society: Literature Review and Future Research Directions." Government Information Quarterly, vol. 39, no. 3, 2022, p. 101708, https://doi.org/10.1016/j.giq.2022.101708.

[9] Goodell, John W., Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis." Journal of Behavioral and Experimental Finance 32 (2021): 100577.

[10] Elkefi, S., & Tounsi, A.," Examining public perceptions and concerns about the impact of heatwaves on health outcomes using Twitter data," The Journal of Climate Change and Health, 17, 100320, 2024, https://doi.org/10.1016/j.joclim.2024.100320

[11] S. Ruchi and P. Srinath, "Big Data Platform for Enterprise project management digitization using Machine learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology, 2018, pp. 1479-1484, doi: 10.1109/ICECA.2018.8474799.

[12] Date, Saroj S., Mahesh B. Shelke, Kiran V. Sonkamble, and Sachin N. Deshmukh. "A Systematic Survey on Text-based Dimensional Sentiment Analysis: Advancements, Challenges, and Future Directions." Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications, 2023

[13] A. F. Hidayatullah, ''Language tweet characteristics of Indonesian citizens,'' in Proc. Int. Conf. Sci. Technol. (TICST), Pathum Thani, Thailand, Nov. 2015, pp. 397–401, doi: 10.1109/TICST.2015.7369393\

[14] Ghosh, Soumitra, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. "Multitasking of sentiment detection and emotion recognition in code-mixed data." Knowledge-Based Systems 260,2023, pp 110182.

[15] A. A. Farzindar and D. Inkpen, "Natural language processing for social media," Synthesis Lectures on Human Language Technologies, vol. 13, no. 2, pp. 1-219, 2020.

[16] A. J. Sietsma, R. W. Groenendijk, and R. Biesbroek, "Progress on climate action: a multilingual machine learning analysis of the global stocktake," Climatic Change, vol. 176, p. 173, 2023. doi: 10.1007/s10584-023-03649-3.

[17] K. Watanabe, "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages," Communication Methods and Measures, vol. 15, no. 2, pp. 81-102, 2021.

[18] Schrader, S. R., and Gultepe, E., "Analyzing Indo-European Language Similarities Using Document Vectors," Informatics, vol. 10, p. 76, 2023. doi: 10.3390/informatics10040076.

[19] Maity, K., Sen, T., Saha, S., and Bhattacharyya, P., "MTBullyGNN: A Graph Neural Network-Based Multitask Framework for Cyberbullying Detection," IEEE Transactions on Computational Social Systems, vol. 11, pp. 849-858, 2024. doi: 10.1109/TCSS.2022.3230974.

[20] Pandey, R., and Singh, J. P., "BERT-LSTM model for sarcasm detection in code-mixed social media posts," Journal of Intelligent Information Systems, vol. 60, pp. 235-254, 2023. doi: 10.1007/s10844-022-00755-z.

[21] Nascimento, Francimaria R., Cavalcanti, G. D., and Da Costa-Abreu, M., "Unintended Bias Evaluation: An Analysis of Hate Speech Detection and Gender Bias Mitigation on Social Media Using Ensemble Learning," Expert Systems with Applications, vol. 201, p. 117032, 2022. doi: 10.1016/j.eswa.2022.117032.
[22] Manikyamba, I. L., and Mohan, A. K., "Using a Novel Hybrid Krill Herd and Bat based Recurrent Replica to Estimate the Sentiment Values of Twitter based Political Data," 2023.
[23] M. Alhawarat and A. O. Aseeri, "A Superior Arabic Text Categorization Deep Model (SATCDM)," in IEEE Access, vol. 8, pp. 24653-24661, 2020, doi: 10.1109/ACCESS.2020.2970504.
[24] Altaf, A., Anwar, M. W., Jamal, M. H., et al., "Exploiting linguistic features for effective sentence-level sentiment analysis in Urdu language," Multimedia Tools and Applications, pp. 1–27, 2023.
[25] Ghasemi, R., Ashrafi Asli, S. A., and Momtazi, S., "Deep Persian sentiment analysis: Cross-lingual training for low-resource languages," Journal of Information Science, vol. 48, pp. 449-462, 2022. doi: 10.1177/0165551520962781.
[26] Thara, S., and Poornachandran, P., "Social media text analytics of Malayalam–English code-mixed using deep learning," Journal of Big Data, vol. 9, pp. 1–25, Dec. 2022.
[27] González Godino, I., and DHaro, L. F., "TASS: Sentiment analysis of tweets for Spanish variants," Proceedings of TASS, 2019.
[28] Neumann, M., and Vu, N. T., "Cross-lingual and multilingual speech emotion recognition on English and French," in Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 5769–5773, 2018. doi: 10.1109/ICASSP.2018.8462162.
[29] Saumya, S., Kumar, A., and Singh, J. P., "Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approach," Engineering Applications of Artificial Intelligence, vol. 133, p. 108159, 2024. doi: 10.1016/j.engappai.2024.108159.
[30] Mandl, T., Modha, S., Kumar, M. A., and Chakravarthi, B. R., "Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English, and German," in Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 29–32, 2020.
[31] Khan, L., Amjad, A., Afaq, K. M., et al., "Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media," Applied Sciences, vol. 12, p. 2694, 2022.
[32] Baruah, A., Das, K. A., Barbhuiya, F. A., and Dey, K., "IIITG-ADBU at SemEval2020 task 9: SVM for sentiment analysis of English-Hindi code-mixed text," in Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain, Dec. 2020. Association for Computational Linguistics, ACM.
[33] Yadav, S., and Chakraborty, T., "Unsupervised sentiment analysis for code-mixed data," arXiv preprint, arXiv:2001.11384, ACL, 2020.
[34] Choudhary, N., Singh, R., Bindlish, I., and Shrivastava, M., "Sentiment analysis of code-mixed languages leveraging resource rich languages," in 19th International Conference on Computational Linguistics and Intelligent Text Processing, Hanoi, Vietnam, 2018.
[35] Banerjee, S., Ghannay, S., Rosset, S., Vilnat, A., and Rosso, P., "Recurrent convolutional neural network for code-mixed sentiment analysis," in Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain, Dec. 2020. Association for Computational Linguistics, ACM.
[36] Shekhawat, S. S., Shringi, S., and Sharma, H., "Twitter sentiment analysis using hybrid Spider Monkey optimization method," Evolutionary Intelligence, vol. 14, pp. 1307–1316, 2021. doi: 10.1007/s12065-019-00334-2.
[37] Dowlagar, S., and Mamidi, R., "CMSAOne@Dravidian-CodeMix-FIRE2020: A Meta Embedding and Transformer model for Code-Mixed Sentiment Analysis on Social Media Text," in FIRE (arXiv), 2020.
[38] Shekhawat, S. S., Shringi, S., and Sharma, H., "Twitter sentiment analysis using hybrid Spider Monkey optimization method," Evolutionary Intelligence, vol. 14, pp. 1307–1316, 2021. doi: 10.1007/s12065-019-00334-2.
[39] Wrigley Kelly, N. E., Murray, K. E., McCarthy, C., and O'Shea, D. B., "An objective analysis of quality and readability of online information on COVID-19," Health Technology (Berl), vol. 11, pp. 1093-1099, 2021. doi: 10.1007/s12553-021-00574-2.

[40] Wang, S., Liu, X., and Zhou, J., "Readability is decreasing in language and linguistics," Scientometrics, vol. 127, pp. 4697-4729, 2022.
[41] Wrigley Kelly, N. E., Murray, K. E., McCarthy, C., and O'Shea, D. B., "An objective analysis of quality and readability of online information on COVID-19," Health Technology (Berl), vol. 11, pp. 1093-1099, 2021. doi: 10.1007/s12553-021-00574-2. Epub 2021 Jun 24. PMID: 34189011; PMCID: PMC8222704.
[42] Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., and Antani, S. K., "Assessing inter-annotator agreement for medical image segmentation," IEEE Access, vol. 11, pp. 21300-21312, 2023.
[43] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., & Shrivastava, M.," A dataset of Hindi–English code-mixed social media text for hate speech detection" In second workshop on computational modelling of people's opinions, personality, and emotions in social media, Association for Computational Linguistics (ACL)
[44] Joshi, A., Prabhu, A., Shrivastava, M., and Varma, V. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text, COLING, pages 2482–2491, ACL Anthology
[45] Santosh, T. Y. S. S., and Aravind, K. V. S., "Hate speech detection in Hindi-English code-mixed social media text," ACM International Conference Proceedings Series, 2019. doi: 10.1145/3297001.3297048.
[46] Wang, S., and Manning, C. D., "Baselines and bigrams: Simple, good sentiment and topic classification," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, pp. 90–94, Stroudsburg, PA, USA, Association for Computational Linguistics, ACM.
[47] Sharma, S., Srinivas, P. Y. K. L., and Balabantaray, R. C., "Text normalization of code mix and sentiment analysis," Advances in Computing, Communications and Informatics, 2015, pp. 1468–1473.
[48] Johnvictor, A. C., Durgamahanthi, V., Pariti Venkata, R. M., and Jethi, N., "Critical review of bio-inspired optimization techniques," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 14, p. e1528, 2022.
[49] Wolpert, D. H., "What is important about the no free lunch theorems?," in Black Box Optimization, Machine Learning, and No-Free Lunch Theorems, Cham: Springer International Publishing, pp. 373-388, 2021.
[50] Trojovská, E., Dehghani, M., and Trojovský, P., "Fennec fox optimization: a new nature-inspired optimization algorithm," IEEE Access, vol. 10, pp. 84417-84443, 2022.
[51] Jordan, M., Kornowski, G., Lin, T., Shamir, O., and Zampetakis, M., "Deterministic nonsmooth nonconvex optimization," in The Thirty Sixth Annual Conference on Learning Theory, pp. 4570-4597, PMLR, 2023.
[52] Lan, G., "First-Order and Stochastic Optimization Methods for Machine Learning," vol. 1, Cham: Springer, 2020.
[53] Boussaïd, I., Lepagnot, J., and Siarry, P., "A survey on optimization metaheuristics," Information Sciences, vol. 237, pp. 82-117, 2013.
[54] Liu, Y., Gao, Y., and Yin, W., "An improved analysis of stochastic gradient descent with momentum," Advances in Neural Information Processing Systems, vol. 33, pp. 18261-18271, 2020.
[55] Akinola, O. O., Ezugwu, A. E., Agushaka, J. O., Zitar, R. A., and Abualigah, L., "Multiclass feature selection with metaheuristic optimization algorithms: a review," Neural Computing and Applications, vol. 34, no. 22, pp. 19751-19790, 2022.
[56] Mao, X., Yuan, K., Hu, Y., Gu, Y., Sayed, A. H., and Yin, W., "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," IEEE Transactions on Signal Processing, vol. 68, pp. 2513-2528, 2020.
[57] Li, J., An, Q., Lei, H., Deng, Q., and Wang, G.-G., "Survey of Lévy flight-based metaheuristics for optimization," Mathematics, vol. 10, no. 15, p. 2785, 2022.
[58] Samanta, B., Ganguly, N., and Chakrabarti, S., "Improved sentiment detection via label transfer from monolingual to synthetic code-switched text," arXiv preprint, arXiv:1906.05725, 2019.
[59] Kamble, S., and Joshi, A., "Hate speech detection from code-mixed Hindi-English tweets using deep learning models," arXiv preprint, arXiv:1811.05145, 2018.
[60] Sharma, A., Kabra, A., and Jain, M., "Ceasing hate with moh: Hate speech detection in Hindi–English code-switched language," Information Processing & Management, vol. 59, p. 102760, Elsevier, 2022.